NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

SPORTS ANALYTICS: MAXIMIZING PRECISION IN PREDICTING MLB BASE HITS

Pedro Miguel Gonçalves André Alceo

Dissertation presented as the partial requirement for obtaining a Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

SPORTS ANALYTICS: MAXIMIZING PRECISION IN PREDICTING MLB BASE HITS

Pedro Miguel Gonçalves André Alceo

Dissertation presented as the partial requirement for obtaining a Master's degree in Information Management, Specialization Knowledge Management and Business Intelligence

Advisor: Roberto André Pereira Henriques

February 2019

ACKNOWLEDGEMENTS

The completion of this master thesis was the culmination of my work accompanied by the endless support of the people that helped me during this journey. This paper would feel emptier without somewhat expressing my gratitude towards those who were always by my side.

Firstly, I would like to thank my supervisor Roberto Henriques for helping me find my passion for data mining and for giving me the guidelines necessary to finish this paper. I would not have chosen this area if not for you teaching data mining during my master's program.

To my parents João and Licínia for always supporting my ideas, being patient and for along my life giving the means to be where I am right now. My sister Rita, her husband João and my nephew Afonso, whose presence was always felt and helped providing good energies throughout this campaign.

To my training partner Rodrigo who was always a great company, sometimes an inspiration but in the end always a great friend. My colleagues Bento, Marta and Rita who made my master's journey unique and very enjoyable. To my closest friends who have accompanied me throughout my life that are too many to name but have always provided a great environment for companionship and self-growth.

RESUMO

Nos últimos anos o mundo do desporto alcançou níveis de crescimento nunca antes visto e, este evento, fomentou a necessidade para o crescimento no uso de ferramentas que tragam vantagens para as organizações e os respetivos stakeholders. Como resultado tem se registado um rápido crescimento no uso da análise de dados para vários tópicos relacionados com o desporto que consequentemente origina melhores e rápido julgamentos para os tomadores de decisão.

Nesta linha de pensamento, o principal objetivo deste projeto é contruir um modelo preditivo capaz de prever as probabilidades de um jogador da MLB obter um "base hit" num dia com o propósito de ganhar o jogo *Beat the Streak* e, ao mesmo tempo, providenciar informações valiosas à equipa técnica.

A arquitetura que serviu de diretriz a este projeto foi o CRIPS-DM, o qual foi aplicado a uma base de dados construída especificamente para este projeto com dados publicamente acessíveis. Para alcançar os referidos objetivos, foram usados o Excel com meio para recolher e estruturar a base de dados e o Python para os restantes processos com um enfase na biblioteca *SKlearn.* Os elementos que separam as construções dos modelos finais foram o balanceamento da base de dados, *outliers*, redução da dimensionalidade, seleção das variáveis e os algoritmos – *Logistic Regression, Multi-layer Perceptron, Random Forest* e *Stochastic Gradient Descent.*

Os resultados obtidos foram positivos sendo o modelo com a melhor performance um Multi-layer Perceptron que obteve 85% de escolhas certas no set de teste. Este resultado alcançou uma melhoria de 5 pontos percentuais sobre o melhor modelo encontrado durante a pesquisa bibliográfica.

Os resultados em questão foram positivos, mas existe margem para melhorar os modelos desenvolvidos ou a criação de outros modelos porque com os resultados obtidos ainda é difícil ganhar o jogo Beat the Streak, o que deixa em aberto a possibilidade para a criação de novos modelos.

PALAVRAS-CHAVE

Machine Learning; Data Mining; Análise Predictiva; Modelos de Classificação; Baseball; MLB.

ABSTRACT

As the world of sports expanded to never seen levels, so did the necessity for tools which provided material advantages for organizations and other stakeholders. This resulted in an increase on the use of data and analytics for a multitude of sports related topics, which led to more precise and quicker judgements for decision makers related to sports.

In this line of though, the main objective of this paper is to build a predictive model capable of predicting what are the odds of a baseball player getting a base hit on a given day, with the intention of both winning the game Beat the Streak and to provide valuable information for the coaching staff.

CRISP-DM was the architecture chosen as the main guideline to apply on the dataset, entirely built for this paper, using publicly available data. To achieve these objectives, Excel was used for data collection purposes and Python for the remaining steps with a big emphasis on the SKlearn library. Several models were tested and the main constrains that separate them from each other are balancing, outliers, dimensionality reduction, variable selection and the type of algorithm – Logistic Regression, Multi-layer Perceptron, Random Forest and Stochastic Gradient Descent.

The results obtained were positive, in which one of the Multi-layer Perceptron achieved an 85% correct pick ratio on the test set, which is an improvement of 5 percentage points over the best model found during the literature review.

Nevertheless, there is undoubtedly room for improvements in the final models and for other models with similar intentions, since the results achieved do not provide a good change of Beating the Streak.

KEYWORDS

Machine Learning; Data Mining; Predictive Analysis; Classification Model; Baseball; MLB.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS AND ACRONYMS

2B	Double
AB	At-Bat
AUC	Area Under Curve
BB	Base-on-Balls
CRIPS-DM	Cross Industry Standard Processes for Data Mining
H	Hit
HBP	Hit by Pitch
HIP	Hit in Play
HR	Home Run
KDD	Knowledge Discovery in Databases
LG	Logistic Regression
MLB	Major League Baseball
MLP	Multi-Layer Perceptron
NBA	National Basketball Association
NFL	National Football League
NHL	National Hockey League
OBP	On Base Percentage
OIP	Out in Play
PA	Plate Appearance
PCA	Principle Component Analysis
RC	Runs Created
RF	Random Forest
RFE	Recursive Feature Elimination
SEMMA	Sample, Explore, Modify, Model and Assess
SGD	Steep Gradient Descent
SO	Strikeout

1. INTRODUCTION

"I'm sort of a baseball agnostic; I make it a point never to believe anything just because it is widely known to be so."

— Bill James[1]

1.1. CONTEXT AND RELEVANCE

In the past few years the professional sports market has been growing impressively. Events such as the Super Bowl, the Summer Olympics and the UEFA Champions League are fine examples of the dimension and global interest that can be generated by this industry currently. As the stakes grow bigger and further money and other benefits are involved in the market, new technologies and methods surge to improve stakeholder success (Mordor Intelligence, 2018).

The particular advancement most relevant for this project was the explosion of data creation and data storage systems, during the XXI century, which led to volumes of information that have never been so readily at our disposal before (Cavanillas, Curry, & Wahlster, 2016).Consequently, sports as many other industries could now use data to their advantage in their search for victory and, thus the sports analytics began its ascension to the mainstream (Gera et. All, 2016).

For most organizations winning is the key factor for good financial performance since it provides return in the form of fan attendance, merchandising, league revenue and new sponsorship opportunities (Collignon & Sultan, 2014). Sports analytics is a mean to reach this objective, by helping coaches, scouts, players and other personnel making better decisions based on data, leading to short and long-term benefits for stakeholders of the organization (Alamar, 2013).

The growing popularity of sports and the widespread of information also resulted in the growth of betting in sports events. This resulted in a growth of sports analytics for individuals outside sports organizations, as betting websites started using information based analytical models to refine their odds and gamblers to improve their earnings (Mann, 2018).

Finally, according to a study from Mordor Intelligence (2018) the sport who currently takes the most out of sports analytics is baseball. This is partially the consequence of historical events such as Moneyball where the use of analytics proved to have great effects in the outcome of the Oakland Athletics season, a baseball team which had the least money to spend on players in the League. For most, Moneyball was the turning point in analytics and baseball, which opened the way for the use of analytics in both baseball and other sports (Lewis, 2004).

1.2. PROBLEM

According to the MLB.com Glossary (MLB, 2018a), the definition of a base hit is:

"A hit occurs when a batter strikes the baseball into fair territory and reaches base without doing so via an error or a fielder's choice."

[1] Bill James is an American baseball writer, historian and statistician (for more info go to https://www.billjamesonline.com)

A base hit is a common way for a player to reach a base and to advance other team mates already on base to potentially score runs. Therefore, base hits are among the best outcomes for batter, during his at-bat.

As easy as it sounds, batting the ball properly into the field of play is a big challenge even for professionals and, for examples, by taking the best batters from the 3 previous regular seasons (2015, 2016, 2017), in terms of hit% ([Hits/Plate Appearances]*100), we get Dee Gordon with 31,4%, Daniel Murphy with 31,6% and Jose Altuve with 30,8% respectively (Baseball Reference, 2018). In other words, for every attempt these batters had, they achieved a base hit only 31% of times. According to the dataset built for this paper around 66% of players (not accounting for pitchers batting) get at least one base hit during a game, which on average comprises 4 attempts per game (or plate appearances).

Additionally, during the 4 seasons comprised in the database or 9.720 games played the average streak was 2 and the 2 longest hitting streak achieved, allowing inter season streaks, was 28 by Jackie Bradley Jr. and 30 by Freddie Freeman. These reveals that picking the same player over and over may also not be a viable strategy.

For instance, to win the MLB Beat the Streak[2] it is required that the participant correctly picks 56 times in row a player who gets a base hit in a given day. There are two important rules that should be considered. The first is that the maximum number of players one could pick in a single day is two, and the other is that it is not mandatory to make a pick every single day to keep the streak.

The main problem arises here, which in other words means that random guess does not provide a fair chance to win this game, which translated into a probability:

$$P(Win\ beat\ the\ Streak\ Randomly) = 0{,}66^{56} = 0{,}000000000078$$

Equation 1: Probability of winning MLB beat the streak
Source: Made by the author

1.3. OBJECTIVE

The objective of this project is to build a database and a data mining model capable of predicting which MLB batters are most likely to get a base hit on a given day. In the end, the output of the work can have one of two uses:

1. To give a methodical approach for coach's decision making and on what players should have an advantage on a given game and therefore make the starting lineup;
2. To improve one's probabilities of winning the game MLB Beat the Streak**.**

For the construction of the database, it was collected data from the last four regular seasons of the MLB, from open sources. Regarding the granularity of the dataset, a sample consists on the variables of a player in a game. Additionally, in the dataset it was not considered pitchers batting nor players who had less than 3 plate appearances in the game. Finally, the main categories of the variables used in the models are:

[2] https://www.mlb.com/apps/beat-the-streak

- Batter Performance
- Batter's Team Performance
- Opponent's Starting Pitcher Performance
- Opponent's Bullpen Performance
- Weather
- Ballpark

Regarding the batter's performance those variables will include values from Statcast which is "a state-of-the-art tracking technology that allows for the collection and analysis of a massive amount of baseball data (…)" (MLB, 2018b). These stats in combination with a precision driven approach are what really set this paper apart from others from this area. The final output of the project should be a predictive model, which adequately fits the data using one or multiple algorithms (Ensemble) and methods to achieve the most precision, measured day-by-day. The results will then be compared with other similar projections and predictions to measure the success of the approach.

1.4. STUDY OUTLINE

The first chapter focused on exposing the main objectives and ideas of this paper and provide an idea on its present relevance.

The second chapter focus on presenting the literature review for data mining and machine learning architectures used as guidelines during this project, as well as, the evolution of sports analytics, baseball analytics and the state of art of the topic being approach.

Chapter number three depicts the different steps that were performed to reach the conclusions of the project. A top-down overview of the project can be seen here, where it is explicit the various processes from the data collection and dataset creation until the metrics chosen for model evaluation. Throughout the chapter can be found the different decisions and their reasoning for relevant topics, such as data preprocessing or data transformations.

Across the fourth chapter are depicted detailed results from the application of the processes described in the methodology and their analysis. In this chapter, the best models are examined giving insight on what the best hypothesis were, as well as, the best variables for the problem in question. Finally, the best models are compared with different strategies found during the literature review.

The final chapters are the conclusions and limitations, the former focus on providing a concise summary of the work done and the most important insights obtained during the project. The latter exposes what were the biggest barriers and possible improvements that could be done to enhance the performance of the models.

Overall, the results achieved were positive, with a 5-percentage point improvement over other similar projects. Nevertheless, the expected correct pick ratio achieved of 85%, with a multi-layer perceptron, does not offer an optimal probability for Beating the Streak.

2. LITERATURE REVIEW

2.1. DATA MINING

According to Peter Ffoulkers (2017) the amount of raw data available to industries as increased drastically in recent times. More precisely, there is a broad agreement that the size of the digital universe will double every two years at least, or a 50-fold growth from 2010 to 2020. However, the cheer presence of available does not translate into business value nor competitive advantage if not operated correctly. This may lead to the problem that there might be too much data available for organizations, which consequently may prevent the effective use of the data and making it difficult to reach an optimal status of business value creation (Lavalle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011).

The solution for these problems is in the use of DM techniques. DM represents the application of algorithms to extract useful patterns and insight from data and consequently transforming into information and knowledge (Fayyad, Piatetsky-Shapiro & Smyth, 1996). The two main uses for DM are to forecast- using predictive modelling, and to describe- using descriptive modelling (Wang, 2009). The former focus on recognizing the design and relationships in the data and discovering its properties. Whereas, the latter utilizes pre-labelled data to make authoritative predictions about the future using business forecasting and simulations (Agyapong, Hayfron-Acquah & Asante, 2016).

Both descriptive and predictive modelling take great advantage of Machine Learning (ML) techniques to boost the efficiency and effectiveness of DM projects. According to SAS (2018), “ML is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention”. Belcastro, Marozzo, Talia, & Trunfio (2016) claim that by using ML techniques in conjunction with the right DM tools it is possible to perceive more complex phenomena and solve a wider range of problems.

The growth of data and consequent use of DM in several areas led to a need for a model which helped optimizing and standardize the process of insight retrieval from databases. The three most relevant options proposed were CRISP-DM, KDD and SEMMA (Shafique & Qaiser, 2004). As for CRIPS-DM, it was proposed by SPSS, NCR and Daimler Chrysler in 1996. The development of CRIPS-DM led to the publication of CRISP-DM 1.0 in 2000, where the main guidelines were settled for data mining projects using the model.

CRISP-DM provides a structured approach together with guidelines to help an individual execute a data mining project. The CRISP-DM methodology has an iterative nature and consists of six key phases (Pete Chapman et al. 2000):

1. **Business Understanding** – uncover important factors including success criteria, business and data mining objectives and requirements as well as business terminologies and technical terms.
2. **Data Understanding** – data collection, checking data quality and exploring the data to get insight of data to form hypotheses.

3. **Data Preparation** – selection and preparation of the final data set. Includes tasks such as, data cleansing, integration, transformation and variable selection.
4. **Modeling** – selection and application of various data mining models and algorithms.
5. **Evaluation** – interpretation of the models and algorithms used and evaluating whether they achieve the objectives properly or not.
6. **Deployment** – determining the use the obtained results have by organizing, reporting and presenting the gained knowledge when and where needed.

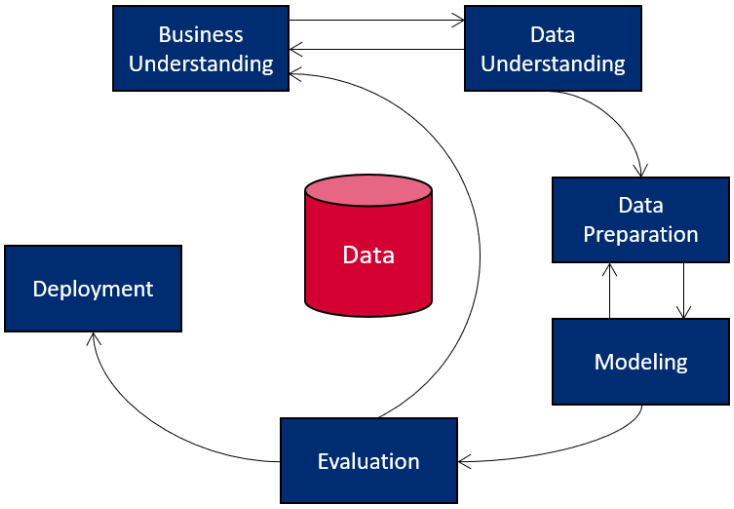


Figure 1 – CRISP-DM architecture
Source: Made by the author, adapted from (Pete Chapman et al. 2000)

2.2. Sports Analytics

Sports and analytics have always had a close relation as in most sports both players and teams are measured by some form of statistics, which are used to provide rankings for both players and teams. During the 20th century, sports started to appeal to the masses and the broadcast of sports events became accessible to the general audience, which led broadcasting companies to start using different statistics to offer a better experience to the audience. The increase in popularity was met with economic benefits for the sports world and researchers started using basic statistics to better understand the games and provide insights to the stakeholders (Albert and Koning, 2008).

2.2.1. Evolution of Sports Analytics

This chapter serves as a bridge which will connect the surge of sports statistics in the mid-20th century to the present. The main driver of this portion will be the identification and understanding of the main studies and events that led to the present concept of sports analytics.

In 1968, Charles Reep was the first data analyst connected to football and developed the base for football notation, which helps categorize each play in a match. Together with the statistician Bernard Benjamin, he looked for insights in 15 years' worth of matches using his notation. The study helped the development of football tactics, suggesting that a more direct style of football would be desirable (Reep & Benjamin, 1968).

In 1977, Bill James was one of the pioneers in the application of statistics to several aspects of baseball. He defied traditional perceptions on how to evaluate players and highlighted the importance of creating runs versus basic statistics like hitting average and earned run average. His ideas captivated a lot of interest and he wrote numerous editions of The Baseball Abstract[3] where he presents advanced statistics and methods that are now considered the foundations for modern sabermetrics[4] (James, 2001).

During the 1980's, Dean Oliver inspired by Bill James sabermetrics began developing analysis on basketball players performance and their contribution to the team. His research and commitment originated what is now known as APBRmetrics[5]. Due to his great developments and achievements, in 2004 he was hired as the first full-time statistical analyst in the NBA (Oliver, 2004).

Even with the success of Reep, Benjamin, Oliver and James sports analytics never settled inside sports organizations until the recent event commonly known as Moneyball. Billy Beane and Peter Brand used several sabermetrics and other analytics tools to create a roster of players which were considerate not very good for most teams and reached the playoffs in one of the most inspiring seasons in MLB history (Lewis, 2004). This event was for most the turning point for the use of analytics in sports.

A recent example of sports analytics taking a team to the next level is the history of the Houston Astros road to win over the Los Angeles Dodgers in the 2017 World Series[6]. The Houston Astros had consecutive losing records from 2011 to 2014 where they traded away their star players and veterans for future benefits, also known as tanking[7], which with the use of a data driven mentally led them to the top of Major League Baseball (Sheinin, 2017).

2.2.2. Data Mining in Sports

Recently, the sports industry is experiencing a growth in terms of data mining models which search for a deeper understanding regarding various aspects of in field and off the field events. These models are being built for a multitude of sports and cover different parts of the game, player's performance and other factors that may affect the outcome of the game (Albert and Koning, 2008).

Most models attempt to predict future events in the sports world, whether for performance or gambling. For example, Edelmann-Nusser, Hohmann and Henneberg (2002) tried to predict the competitive performance of a swimmer at the Summer Olympics Games in 2002 using an artificial neural network. Another example would be the application of decision trees for identifying characteristics in matchups between players and their interactions which drive the result of hockey games (Morgan, Williams & Barnes, 2013). Finally, another popular use of predictive modeling is the forecast of the NCAA men's basketball playoffs, which attracts multiple people to Kaggle contests and fantasy sports websites. In 2015, Yuan et al. present a famous mixture of modelers approach to try and forecast the 2014 version of these playoffs.

[3] http://baseballanalysts.com/archives/2004/07/abstracts_from_12.php

[4] http://sabr.org/sabermetrics

[5] http://www.apbr.org

[6] The World Series is the annual championship series of Major League Baseball (MLB) in North America

[7] In sports, tanking is the mentality of selling/losing present assets to achieve greater benefits in the future

A major factor in understanding how important sports analytics is for a sport and in what areas it is most useful is luck. Mauboussin (2012) tries to calculate the amount of luck present in several sports by giving the assumption that the observed variance in the win-lost record of a regular season is given by the variance of the skill plus the variance of luck for the given sport. In the end of a season it Is known the observer variance and, by simulating the entire season with the results at random he calculates the variance of luck:

$$Var(Observed) = Var(Skill) + Var(Luck)$$

Equation 2: Probability of winning MLB beat the streak
Source: Made by the author, adapted from (Mauboussin, M., 2012)

Finally, he tested the following games over 5 seasons and achieved the following results:

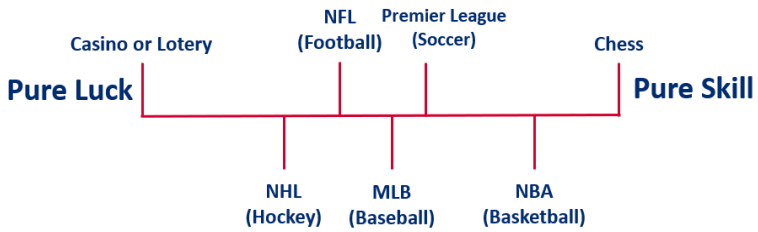


Figure 2: Luck in sports
Source: Made by the author, adapted from (Mauboussin, M., 2012)

According to Mauboussin (2012) these results reflect several aspects of the sports. For example, sports where there are more players in play are harder to predict than sports than sports where a small number of players influence the outcome of the game. Other important factors mentioned are the number of the regular season games, the number of opportunities per game and the type of events that form the game. The latter aspect means that a sport with more discrete events, like baseball, is easier to demonstrate the skill of players than a more fluid type game like hockey. Finally, sports that are bound by other external factors, such as meteorological, altitude or even the distance a team needs to travel to play the game tend to steer to the luck side of the diagram.

Analyzing baseball through this perspective there are mix of the above-mentioned factors that put it around the middle of the diagram. Factors like the number of games are not particularly important for the approach used in this paper since the objective is to understand bases hits instead of season wins. Nevertheless, it is quite interesting that baseball events are among the most discrete in sports, there are not a lot of participants in each event (usually only the batter, the pitcher and one or two fielders) and the number of opportunities to score are arguably high. Even considering the above-mentioned characteristics baseball is still quite random when it comes to base hits and it is crucial to interpret every possible variable, both external and internal, to better predict these outcomes (Bailey, S., 2017).

2.3. Baseball Analytics

The use of analytics in baseball is nowadays a common practice and a lot of historical baseball data is publicly available. According to a study carried by Morton Intelligence (2018) there are more MLB organizations using analytics than in any other major league in North America.

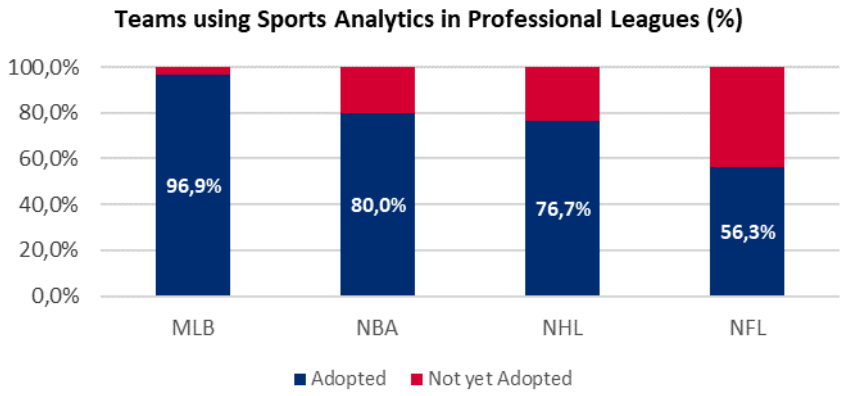


Figure 3: Percentage of Sports organizations that use analytics, in North American Major Leagues
Source: Made by the author, adapted from (Mordor Intelligence, 2018)

2.3.1. State of Data Mining in Baseball

Nowadays, most baseball studies using data mining tools focus on the financial aspects and profitability of the game (Sykora, Chung, Folland, Halkon, & Edirisinghe, 2015). The understanding of baseball in game events is often relegated to sabermetrics: "the science of learning about baseball through objective evidence" (Wolf, 2015). Most sabermetrics studies concentrate in understanding the value of an individual and once again are mainly used for commercial and organizational purposes (Ockerman & Nabity, 2014; Robinson, 2014). The reason behind the emphasis on the commercial side of the sport is that "it is a general agreement that predicting game outcomes is one of the most difficult problems on this field" (Valero, C., 2016) and operating data mining projects with good results often requires investments that demand financial return.

Apart from the financial aspects of the game, predictive modelling is often used to try and predict the outcome of matches (which team wins a game or the number of wins a team achieves in a season) and predicting player's performance. The popularity of this practice grew due to the expansion of sports betting all around the world (Stekler, Sendor, & Verlander, 2010). The results of these models are often compared with the Las Vegas betting predictions, which are used as benchmarks for performance. Projects like these are used to increase ones earning in betting but could additionally bring insights regarding various aspects of the game. (Jia, Wong & Zeng, 2013; Valero, C., 2016).

In conclusion, a big reason baseball models do not reach greater predictive results is due to luck. This concept is explored by Albert (2015) and Albert (2016) where the author creates methods for predicting player's batting average where he emphasizes that around half of the variability of a player batting average can be attributed to luck. In other words, there are several aspects of the game that are hard to translate into data and result in a higher unpredictability in these types of events. Hence the real objective of any data mining model of this type should be to minimize the effect of luck in the model (Bailey, 2017).

2.3.2. Statcast

Statcast is a relatively new data source that was implemented in 2015 across all MLB parks. According to MLB.com Glossary (MLB, 2018b) "Statcast is a state-of-the-art tracking technology that allows for the collection and analysis of a massive amount of baseball data, in ways that were never possible in the past. (…) Statcast is a combination of two different tracking systems -- a Trackman Doppler radar and high definition Chyron Hego cameras. The radar, installed in each ballpark in an elevated position behind home plate, is responsible for tracking everything related to the baseball at 20,000 frames per second. This radar captures pitch speed, spin rate, pitch movement, exit velocity, launch angle, batted ball distance, arm strength, and more."

Prior to Statcast the public had access to data through PITCHf/x, which measured several parameters, including pitch speed, trajectory speed and release point. PITCHf/x was created by Sportvision and was implemented since 2008 in every MLB stadium (Fast, 2010). The video monitoring, provided by Statcast, provides the public access to a wider range of variables, player and ball tracking which is a great breakthrough for both baseball analysts and baseball in general.

(Albert, et al., 2018) developed an independent report using Statcast data to analyze possible causes for the recent surge in Home Runs in the MLB. They used variables such as the launch angle, exit velocity and many others and reached the conclusion that this increase was primarily related to a reduction in drag of the baseballs. This is a great showing of the potential of Statcast and that it is rapidly surpassing the previous methods, like PITCHf/x and could consequently lead to more precise measurements of player's abilities (Sievert & Mills, 2016).

2.3.3. Beat the Streak

The MLB Beat the Streak is a betting game based on the commonly used term hot streak, which in baseball is applied for players that have been performing well in recent games or that have achieved base hits on multiple consecutive games. The objective of the game is to pick 57 times correctly in a row a batter that achieves a base hit on the day that it was picked. The game is called Beat the Streak since the longest hit streak achieved was 56 by the hall of famer Joe DiMaggio, during the 1941 season. The winner of the contest, which in run annually wins US$ 5.600.000, with other prizes being distributed every time a better reaches a multiple of 5 in is streak, for example picking 10 times or 25 times in a row correctly (Beat the Streak, 2018).

Some relevant rules that are important for the strategy of reaching higher streaks are: the better can select 1 or 2 batters per day but note that the streak does not end if no batter is picked for a given day. If the player selected does not start the game for any reason the player is not accounted as an actual pick. Nevertheless, if the player is switched mid game without achieving a base hit, the streak is reset. Finally, there is a Mulligan which works as a second change for when the streak of a better lies between 10 and 15. If the better incorrectly picks during this state of his streak, his streak will remain (Beat the Streak, 2018).

To better visualize the different rules, table 1 illustrates some examples on how the streak works in different scenarios:

Pick 1	Pick 2	Result
Hit	Hit	Streak increases by two (2)
Not Hit	Not Hit	Streak ends and resets to zero unless a Mulligan applies in which case the streak is preserved at the current level
Hit	Not Hit	Streak ends and resets to zero unless a Mulligan applies in which case the streak is preserved at the current level
Pass	Not Hit	Streak ends and resets to zero unless a Mulligan applies in which case the streak is preserved at the current level
Hit	Pass	Streak is increased by one (1)
Pass	Pass	Streak is preserved at the current level

Table 1: Beat the Streak scenario outcomes
Source: Made by the author, adapted from (Beat the Streak, 2018)

2.3.4. Predicting batting performance

Baseball is a game played by two teams who take turns batting (offense) and fielding (defense). The objective of the offense is to bat the ball in play and score runs by running the bases, whilst the defense tries to prevent the offense from scoring runs. The game proceeds with a player on the fielding team as the pitcher, throwing a ball which the player on the batting team tries to hit with a bat. When a player completes his turn batting he gets credited with a plate appearance, which can have one of the following outcomes, as seen below:

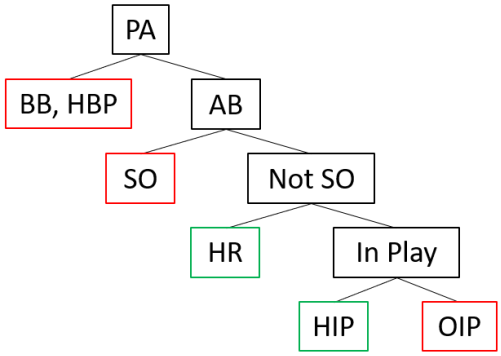


Figure 4: Breakdown of a plate appearance
Source: Made by the author, adapted from (Albert, 2015)

Denotated in green are the events which result on a base hit and in red the events which are not. Thereafter, what this paper tries to achieve is to predict if a batter will achieve a Home Run (HR) or a Ball hit in play (HIP) among all his plate appearances during a game. The most common approach which mostly resembles the model built in this project is forecasting the batting average (AVG). The differences from both approaches are that the batting average does not account for Base on Balls (BB) and Hit by Pitches (HBP) scenarios.

$$Batting\ Average\ (AVG) = \frac{HR+HIP}{AB}$$

$$Hitting\ Percentage\ (H\%) = \frac{HR + HIP}{PA}$$

Equation 3: Batting Average and Hitting Percentage calculation
Source: Made by the author

There are many systems which predict offensive player performance including batting averages. These models range from simple to complex. Henry Druschel from Beyond the Boxscore identifies that the main systems in place are: Marcel[8], PECOTA[9], Steamer[10], and ZiPS[11] (Druschel ,2016; Bailey, 2017).

- Marcel encompasses data from the last three seasons and gives extra weight to the most recent seasons. Then it shrinks a player's prediction to the league average adjusted to the age using a regression towards the mean. The values used for this are usually arbitrary.

- PECOTA uses data for each player using their past performances, with more recent years weighted more heavily. PECOTA then uses this baseline along with the player's body type, position, and age, to identify various comparison players. Using an algorithm resembling k-nearest neighbors, it identifies the closest player to the projected player and the closer this comparison is the more weight that comparison player's career carries.

- Steamer uses a weighted average of past season performances adjusted to the league average. Steamer, much like Marcel, then looks to regress the prediction towards the mean but the degree and weight is regressed varies. Those are set using regression analysis of past players.

- ZiPS like Marcel and Steamer uses a weighted regression analysis but specifically four years of data for experienced players and three years for newer players or players reaching the end of their careers. It then pools players together based on similar characteristics.

Bailey, Loeppky and Swartz (2017) use PECOTA in conjunction with Statcast data to eliminate some of the effect of luck in predicting player's batting average. The objective of the paper is to improve the performance of PECOTA using the aforementioned data. The solution is achieved by simply combining both techniques using a linear regression. The results are a very marginal improvement that prove that there are potential gains in using Statcast data and variables that come from this source.

Goodman and Frey (2013), developed a machine learning model to predict the batter most likely to get a hit each day. Their objective was to win the MLB Beat the Streak game, to do this they built a generalized linear model (GLM) based on every game since 1981. The results on testing were 70% precision on correct picks and in a real-life test achieved a 14-game streak with a peak precision of 67,78%

Clavelli and Gottsegen (2013), created a data mining model with the objective of maximizing the precision of hit predictions in baseball. The project compiles game data from previous seasons and uses a logistic regression, which achieves a 79,3% precision on its testing set. In the paper it is also used a support vector machine, which heavily overfitted and only achieved a 63% precision in its testing set.

[8] Marcel - Available at http://www.tangotiger.net/marce
[9] PECOTA - Available at https://www.baseballprospectus.
[10] Steamer - Available at http://steamerprojections.com/blog/about-2
[11] ZiPS – Available at https://www.fangraphs.com

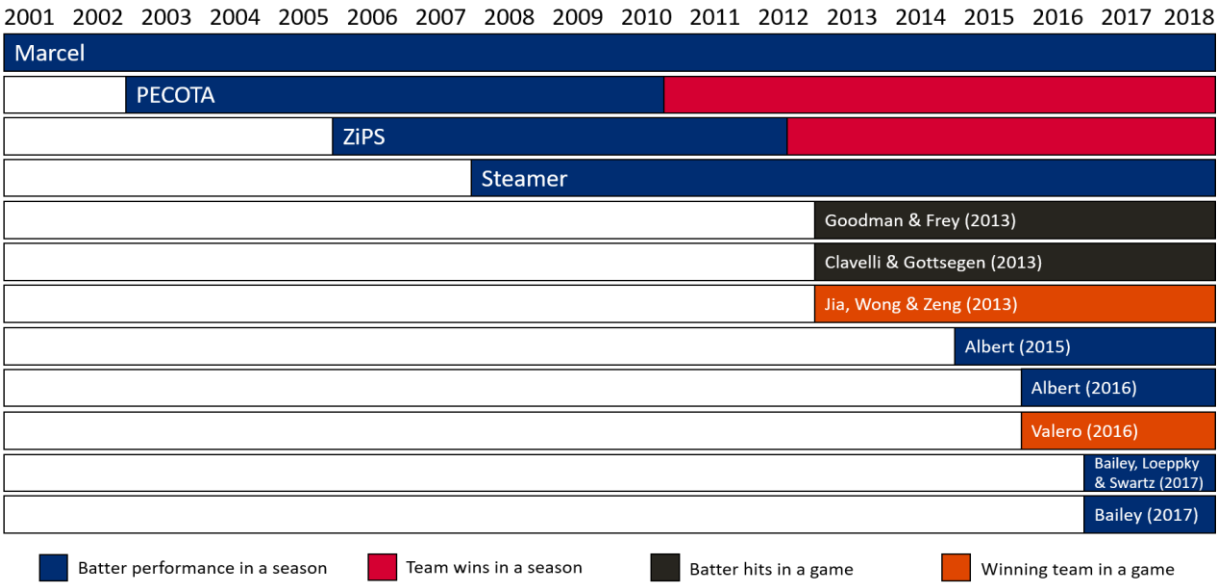


Figure 5: Literature review summary
Source: Made by the author

3. METHODOLOGY

This section of the paper presents the methodological processes done from the collection and creation of the dataset until the evaluation of the final models. As explored during the literature review, the project follows the popular architecture for data mining projects CRISP-DM and, this chapter is focused on the data understanding, data preparation, modelling and evaluation processes.

Figure 6 depicts the various software used to complete these tasks, which were Microsoft Excel and Python. In a first stage, Microsoft Excel was used for data collection, data integration and for variable transformation purposes. In a second stage, the dataset was imported to Python where the remaining data preparation, modeling and evaluation processes were carried out. In Python the three crucial packages used were Pandas (dataset structure and data preparation), Seaborn (data visualization) and Sklearn (for modeling and model evaluation).

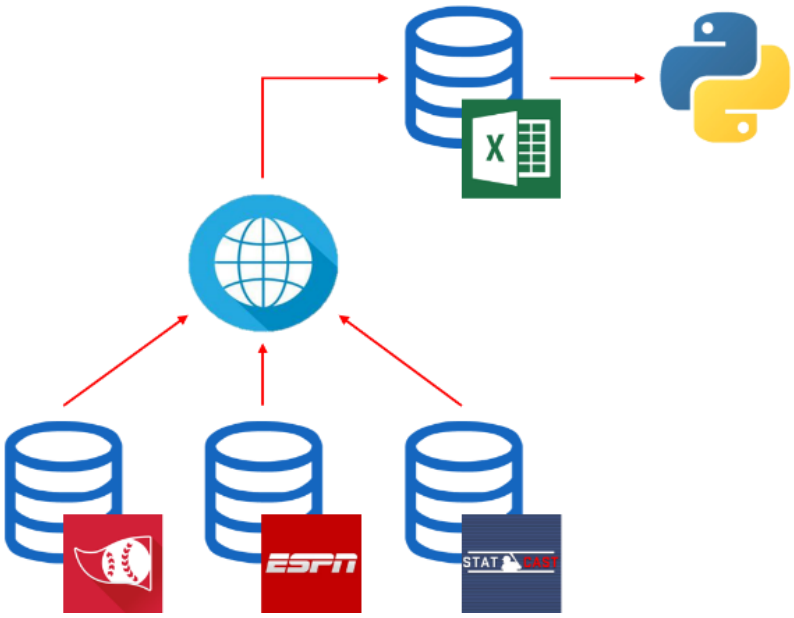


Figure 6: Data sources and data management diagram
Source: Made by the author

Scikit-learn is a popular Python library that has a collection of machine learning algorithms, for both supervised and unsupervised learning. Furthermore, it includes many functions and applications that enable some preprocessing, modelling and evaluation steps of a data mining project. Finally, it is known for being easy to integrate in many applications since it relies on the Python ecosystem and thus is used in wide range of subjects (Michel et al., 2011, Pedregosa et al., 2011)

Finally, figure 7 illustrates an overall view of the project, showing the process performed on each of the training sets and the main constrains that separate the final 48 models, beginning in the raw dataset and finishing in the evaluation procedures, note that the final evaluation procedures were performed for both the validation set, i.e. data partitioning the training set using a 10-fold cross validation and for the test set which was separated initially and, the only treatment perform on the latter was normalization of the variables.

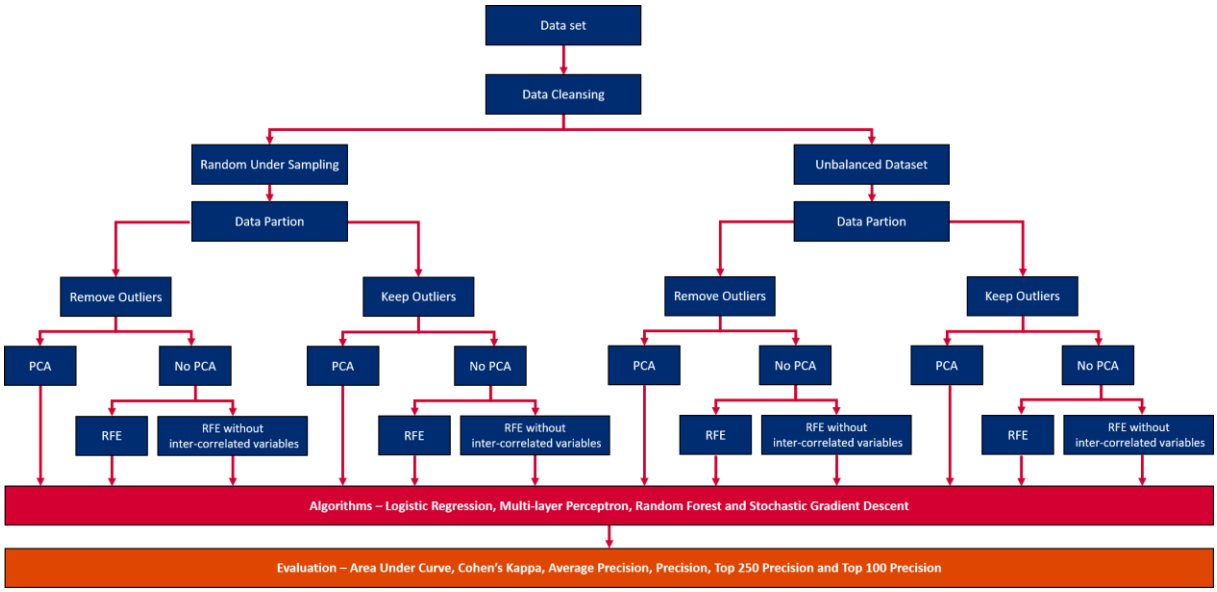


Figure 7: Top down model creation diagram
Source: Made by the author

3.1. DATA COLLECTION

This sub-chapter will describe which data sources were used and resumes the process that was carried out to arrive to the final dataset. The data considered for this study were the games played in Major League Baseball for the seasons 2015, 2016, 2017 and 2018.

The features were collected from the three open source websites:

1. **Baseball Reference** – Baseball Reference is a subsection of the Sports Reference website; the latter includes several other sport related websites. They attempt to give a comprehensive approach to sports data. In their baseball section it is possible to find extensive information about baseballs teams, baseball players, baseball statistics and other baseball related themes dating back to 1871. The data collected from this source is game-by-game player statistics and weather conditions, which could be sub-divided into batting statistics, pitching statistics and team statistics. Data was displayed in a box-score like manner, has seen in the table 2 (Baseball Reference, 2018).
2. **ESPN** – ESPN is a famous North American sports broadcaster with numerous television and radio channels. ESPN mainly focus on covering North American professional and college sports such as, basketball, American football or baseball. Their website contains live scores, news, statistics and other sports related information up to date. The resource retrieved from the ESPN website was the ballpark factor (ESPN, 2018).
3. **Baseball Savant** – Baseball Savant provides player matchups, Statcast metrics and advanced statistics in a simple and easy-to-view way. These include several data visualization applications which help users explore Statcast data. The data retrieved from this data source

includes Statcast yearly player statistics, such as average launch angle, average exit velocity, etc (Baseball Savant, 2018).

Player	Date	Tm	Opp	Rslt	PA	AB	R	H	2B	3B	HR	RBI	BB	IBB	SO	HBP	SH	SF	ROE	GDP	SB	CS	WPA	RE24	aLI	BOP	Pos Summary
Tim Beckham	2018-03-31	BAL	MIN	L 2-6	4	4	1	1	0	0	1	2	0	0	2	0	0	0	0	0	0	0	-0.048	0.913	.518	7	3B
Chris Davis	2018-03-31	BAL	MIN	L 2-6	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.051	-0.874	.400	1	1B
Adam Jones	2018-03-31	BAL	MIN	L 2-6	4	3	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	-0.042	-0.766	.495	4	CF
Manny Machado	2018-03-31	BAL	MIN	L 2-6	4	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0.028	0.274	.390	2	SS
Trey Mancini	2018-03-31	BAL	MIN	L 2-6	4	2	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0.051	0.180	.475	5	LF
Anthony Santander	2018-03-31	BAL	MIN	L 2-6	4	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	-0.029	-0.303	.458	8	DH
Jonathan Schoop	2018-03-31	BAL	MIN	L 2-6	4	4	0	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	-0.056	-0.468	.595	3	2B
Chance Sisco	2018-03-31	BAL	MIN	L 2-6	2	2	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	-0.017	-0.522	.325	9	PH C
Danny Valencia	2018-03-31	BAL	MIN	L 2-6	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.002	0.414	.030	6	PH

Table 2: Sports Reference box-score display
Source: Made by the author (Baseball Reference, 2018)

In order to build the final dataset, the different features were saved in Excel Sheets. From Baseball Reference three sub sections were created: batter box-scores, pitcher box-scores and team box-scores. From Sports Savant batter yearly Statcast statistics. From ESPN the ballpark related aspects. Note that, since the data came from different data sources there was a need to integrate the variables into a single dataset. This process was performed in Microsoft Excel, where with the help of player's name, date or other relevant features it was created the necessary primary keys, these in conjunction with the VLOOKUP function in Microsoft Excel enabled the integration of the data. There were some records that did not coincide or that needed further work, are mentioned in the data preparation chapter.

3.2. DATA UNDERSTANDING

This sub-chapter will serve the purpose of explaining the category of the features present in the dataset, to help understanding the variables used to create the models and the transformations applied to the features. These objectives will be met with using descriptive statistics and data visualization techniques.

3.2.1. Category description

Throughout the literature review several projects and papers were analyzed and, from which we could hypothesize the best categories for this paper. Along this chapter the chosen features categories will be explained for a general understanding on what they are and their potential importance for the models. The table 3 includes a description for all the features in the final dataset.

Category	Variable	Description	Original Variable
Date	Date	Date the game was played	✔
	Double Headers	Number of the game if double header was played	✔
	Year	Year the game was played	✔
	Month	Month the game was played	✔
Batter	Team	ID of the batter's team	✔
	Opponent	ID of the batter's opponent	✔
	Batter ID	Batter's ID	✔
	Batter Name	Batter's Name	✔
	Games Played	Number of games played by the batter	✘
	Batter Hand	Batter's usual batting hand	✔
	Batter Hand (2)	Batter's batting hand in this game	✔
	Matchup (B/P)	Batter's hand and pitchers hand	✘
	Batting Order	Position of the batter in the batting order for the game	✔
	Position	Fielding position of the batter	✔
	Road/Home	Batter played Home or on the Road	✔
	PA	Plate appearances in the game	✔
	AB	At bats in the game	✔
	2B	Doubles in the game	✔
	HR	Home runs in the game	✔
	BB	Base on balls in the game	✔
	SO	Strikeouts in the game	✔
	H% (7games)	Hitting percentage of the batter in the last 7 games played	✘
	H% (15games)	Hitting percentage of the batter in the last 15 games played	✘
	H% (30games)	Hitting percentage of the batter in the last 30 games played	✘
	SO% (7games)	Strikeout percentage of the batter in the last 7 games played	✘
	SO% (15games)	Strikeout percentage of the batter in the last 15 games played	✘
	SO% (30games)	Strikeout percentage of the batter in the last 30 games played	✘
	BB% (7games)	Base on ball percentage of the batter in the last 7 games played	✘
	BB% (15games)	Base on ball percentage of the batter in the last 15 games played	✘
	BB% (30games)	Base on ball percentage of the batter in the last 30 games played	✘
	2B% (7games)	Double percentage of the batter in the last 7 games played	✘
	2B% (15games)	Double percentage of the batter in the last 15 games played	✘
	2B% (30games)	Double percentage of the batter in the last 30 games played	✘
	HR% (7games)	Home run percentage of the batter in the last 7 games played	✘
	HR% (15games)	Home run percentage of the batter in the last 15 games played	✘
	HR% (30games)	Home run percentage of the batter in the last 30 games played	✘
	AB (7games)	Number of at bats in the last of the batter in the last 7 games	✘
	AB (15games)	Number of at bats in the last of the batter in the last 15 games	✘
	AB (30games)	Number of at bats in the last of the batter in the last 30 games	✘
	Hit Streak	Batter's current hitting streak	✘
	Average Launch Angle	Batter's average launch angle (yearly)	✔
	Average Exit Velocity	Batter's average Exit Velocity (yearly)	✔
	Brls/PA %	Percentage Barreled balls per plate apperance (yearly)	✔
	Percentage Shift	Percentage shift was used against the batter (yearly)	✔
	H% vs Pitcher	Batter's hit percentage versus this game starting picther	✘
Batting Team	OBP (7games)	Batter's team on base percentage in the last 7 games	✘
	OBP (15games)	Batter's team on base percentage in the last 15 games	✘
	OBP (30games)	Batter's team on base percentage in the last 30 games	✘
Starting Pitcher	Starter	Starting pitcher name	✔
	Number of Starts	Number of games started by the pitcher	✘
	Throwing Hand	Throwing hand of the starting pitcher	✔
	Hit/Inn (3 games)	Hit per inning allowed by the opponent's pitcher in the last 3 games	✘
	Hit/Inn (5 games)	Hit per inning allowed by the opponent's pitcher in the last 5 games	✘
	Hit/Inn (10 games)	Hit per inning allowed by the opponent's pitcher in the last 10 games	✘
Bullpen	Hit/Inn (3 games)	Hit per inning allowed by the opponent's bullpen in the last 3 games	✘
	Hit/Inn (5 games)	Hit per inning allowed by the opponent's bullpen in the last 5 games	✘
	Hit/Inn (10 games)	Hit per inning allowed by the opponent's bullpen in the last 10 games	✘
Weather	Temperature	Temperature (Fº) at the start of the game	✔
	WindSpd	Wind Speed (MPH) at the start of the game	✔
BallPark	BallPark Name	BallPark name	✔
	Altitude	Altitude of the ballpark	✔
	Roof Type	Roof type of the ballpark	✔
	ESPN Hit Factor	ESPN hit factor for the ballpark	✔
Target Variable	Hit (2)	If the batter got a base hit	✔

Table 3: Variable description
Source: Made by the author

A. Batter's performance

These variables look to describe characteristics, conditions or the performance of the batter. These variables translate into data features like the short/long term performance of the batter, tendencies that might prove beneficial to achieve base hits or even if the hand matchup, between the batter and pitcher, is favorable. The reason behind the creation of this category is that selecting good players based on their raw skills is a worthwhile advantage for the model.

B. Batter's team performance

The only aspect that fits this category is the on base percentage (OBP) relative to the team's batter. Since baseball offense is constituted by a 9-player rotation if the batter's team mates perform well, i.e. get on base, this leads to more opportunities for the batter and consequently higher number of at-bats to get a base hit.

C. Opponent starting pitcher's performance

The variables in this category refer to recent performance of the starting pitcher. These variables relate to the pitcher's performance in the last 3 to 10 games and the number of games played by the starting pitcher. The logic behind the category is that the starting pitcher has a big impact on preventing base hits and the best pitchers tend to allow fewer base hits than weaker ones.

D. Opponent bullpen's performance

This category is quite similar to the previous one. Whereas the former category looks to understand the performance of the starting pitchers the latter focus on the performance of the bullpen, i.e. the remaining pitchers that might enter the game when starting pitcher get injured, get tired or enter to create tactical advantages. The reasoning for this category is exactly the same as the previous one, a weaker bullpen tends to provide a higher change of base hits than a good one.

E. Weather Conditions

In terms on weather conditions, the features that are taken into account are wind speed and temperature. Firstly, the temperature affects a baseball game in 3 main aspects: the baseball physical composition, the player's reactions and movements, and the baseball's flight distance (Koch & Panorska, 2013). If all other aspects remain constant higher temperatures lead to a higher chance of offensive production and thus base hits. Secondly, wind speed affects the trajectory of the baseball, which can lead to lower predictability of the ball's movement and even the amount of time a baseball spends in the air (Chambers, Page & Zaidinis, 2003).

F. Ballpark

Finally, ballpark englobes the ESPN ballpark hit factor, the roof type and the altitude. The "Park factor compares the rate of stats at home vs. the rate of stats on the road. A rate higher than 1.000 favors the hitters. Below favors the pitcher" meaning that this factor will have into consideration several aspects from this or other categories, indirectly (ESPN, 2018). Altitude is another aspect that is crucial to the ballpark, the higher the altitude the ballpark is situated the farther the baseball tends to travel. The previous statement is important to the Denver's Coors Field, widely known for its unusually high offensive production (Kraft & Skeeter, 1995). Finally, the roof type of the ballpark

affects some meteorological metrics, since a closed roof leads to no wind and a more stable temperature, humidity, etc when compared to ballparks with and open roof.

3.2.2. Data Exploration

This chapter serves the purpose of understanding data in a deeper level. Using descriptive statistics and visualization techniques it will be possible to identify certain aspects of the dataset that enable other preprocessing steps, making them more accurate and easier to comprehend.

It should be noted that some of the variables in the dataset were only used for the construction of said dataset and, thus were not explored using the methods mentioned above. Therefore, the variables represented in this chapter are the ones that were taken into consideration for modelling purposes.

3.2.2.1. Descriptive Statistics

Regarding the descriptive statistical analysis of the variables, the results are shown in table 4 for the numeric variables and in table 5 for the nominal variables. For the numerical variables it is possible to access the mean, standard deviation, variance, min-max and the minimum and maximum standard deviation, per feature. The tables enable the understanding of the variables at a univariate level and these values will be particularly useful for outlier detection process.

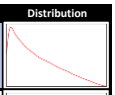
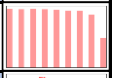
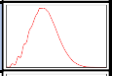
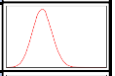
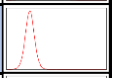
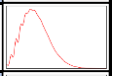
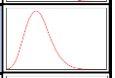
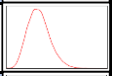
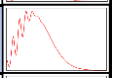
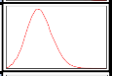
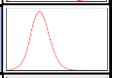
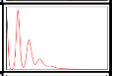
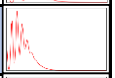
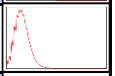
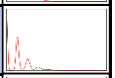
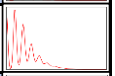
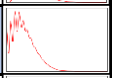
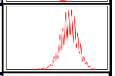
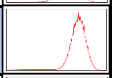
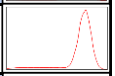
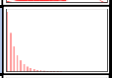
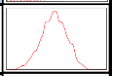
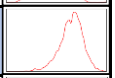
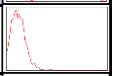
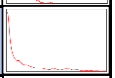
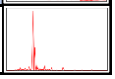
For the categorical values it is represented the number of levels and the mode. These statistics were crucial for variable selection regarding algorithms which cannot handle categorical variables and thus needed some sort of transformation. In sum, variables with a lot of levels are harder to adapt to numeric values if they do not have an order of magnitude and, in these cases are less likely to be picked due to these constrains.

Finally, it was calculated the Pearson's and Spearman's correlation with the objective of uncovering relationships between the independent and with the dependent (target) variable. According to Larose (2005), the former is helpful for not overemphasizing one data component, i.e. using correlated variables might cause the models to become unstable and deliver unreliable results. In contrast, the latter is used for finding features that have a higher predictive power. Therefore, avoid choosing variables that are highly correlated between one another and take special attention to variables that are highly correlated to the target variable.

Pearson's correlation was first described in 1896 and according to Hauke and Kossowski (2011) is" a measure of strength of the relationship between two variables that cannot be measured quantitatively". The main disadvantage of the Pearson's correlation is that it assumes a normal distribution of the features and thrives at finding linear relationships in the data.

With these constrains in mind soon emerged another form of evaluating correlation between features. The Spearman's rank correlation is nonparametric, i.e. does not make assumptions on the distribution of the data and can find nonlinear relationships between the features (Hauke & Kossowki, 2011).

Nevertheless, none of the two is perfect and thus by calculating both types of correlations it will be possible to achieve a better understanding of the relationships between the features and in conclusion make the best use of the data.

Category	Variable	Unit	Average	Std. Dev.	Variance	Minimum	Maximum	Min Std.Dev.	Max. Std.Dev.	Distribution
Batter	Batter Game	Number of Games	182,85	142,53	20.313,69	2,00	634,00	-1,27	3,17	
	Batting Order	1 - 9	4,72	2,46	6,04	1,00	9,00	-1,51	1,74	
	H% (7games)	%	0,23	0,08	0,01	0,00	1,00	-2,80	9,25	
	H% (15games)	%	0,23	0,06	0,00	0,00	1,00	-3,80	12,51	
	H% (30games)	%	0,23	0,05	0,00	0,00	1,00	-4,72	15,49	
	SO% (7games)	%	0,21	0,10	0,01	0,00	1,00	-2,12	8,20	
	SO% (15games)	%	0,21	0,08	0,01	0,00	1,00	-2,56	9,89	
	SO% (30games)	%	0,20	0,07	0,01	0,00	1,00	-2,83	10,98	
	BB% (7games)	%	0,16	0,08	0,01	0,00	0,80	-1,93	7,81	
	BB% (15games)	%	0,16	0,06	0,00	0,00	0,80	-2,80	11,25	
	BB% (30games)	%	0,16	0,05	0,00	0,00	0,80	-3,52	14,13	
	2B% (7games)	%	0,05	0,04	0,00	0,00	0,75	-1,13	17,23	
	2B% (15games)	%	0,05	0,03	0,00	0,00	0,75	-1,56	23,68	
	2B% (30games)	%	0,05	0,02	0,00	0,00	0,75	-1,97	29,80	
	HR% (7games)	%	0,03	0,04	0,00	0,00	0,40	-0,87	10,21	
	HR% (15games)	%	0,03	0,03	0,00	0,00	0,40	-1,15	13,42	
	HR% (30games)	%	0,03	0,02	0,00	0,00	0,40	-1,38	16,06	
	AB (7games)	At Bats	25,89	3,65	13,29	1,00	39,00	-6,83	3,60	
	AB (15games)	At Bats	54,47	9,23	85,13	1,00	76,00	-5,80	2,33	
	AB (30games)	At Bats	105,25	23,99	575,37	1,00	144,00	-4,35	1,62	
	Hit Streak	Number of Games	1,97	2,49	6,21	0,00	30,00	-0,79	11,25	
	Average Launch Angle	Launch Angle (º)	11,57	4,62	21,37	-35,70	41,70	-10,23	6,52	
	Average Exit Velocity	Miles per Hour (MPH)	88,09	3,67	13,45	58,60	115,70	-8,04	7,53	
	Brls/PA %	%	0,05	0,03	0,00	0,00	0,40	-1,67	12,22	
	Percentage Shift	%	0,15	0,21	0,04	0,00	0,94	-0,72	3,83	
	H% vs Pitcher	%	0,23	0,17	0,00	0,00	1,00	-1,40	4,58	

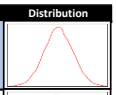
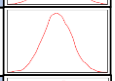
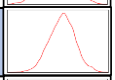
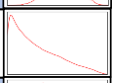
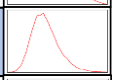
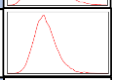
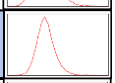
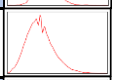
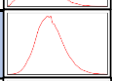
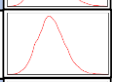
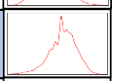
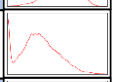
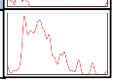
Category	Variable	Unit	Average	Std. Dev.	Variance	Minimum	Maximum	Min Std.Dev.	Max. Std.Dev.	Distribution
Batting Team	OBP (7games)	%	0,32	0,03	0,00	0,20	0,44	-3,54	3,44	
	OBP (15games)	%	0,32	0,02	0,00	0,22	0,43	-3,94	4,67	
	OBP (30games)	%	0,32	0,02	0,00	0,22	0,43	-5,12	6,05	
Starting Pitcher	Starting Pitcher Game	Number of Games	37,91	29,43	866,39	2,00	131,00	-1,22	3,16	
	Hit/Inn (3 games)	Hits per Inning	1,04	0,41	0,17	0,00	30,00	-2,56	71,02	
	Hit/Inn (5 games)	Hits per Inning	1,03	0,35	0,12	0,00	30,00	-2,95	82,93	
	Hit/Inn (10 games)	Hits per Inning	1,02	0,31	0,10	0,00	30,00	-3,30	93,93	
Bullpen	Hit/Inn (3 games)	Hits per Inning	0,97	0,41	0,17	0,00	4,34	-2,39	8,25	
	Hit/Inn (5 games)	Hits per Inning	0,97	0,31	0,10	0,00	2,54	-3,12	5,08	
	Hit/Inn (10 games)	Hits per Inning	0,97	0,22	0,05	0,24	1,91	-3,31	4,28	
Weather	Temperature	Fahrenheit (Fº)	73,70	10,56	111,48	27,00	108,00	-4,42	3,25	
	WindSpd	Miles per Hour (MPH)	7,49	5,04	25,40	0,00	28,00	-1,49	4,07	
Ballpark	Altitude	Feet (ft)	504,80	917,08	841.036,27	0,00	5.197,00	-0,55	5,12	
	ESPN Hit Factor	-	1,00	0,08	0,01	0,84	1,30	-1,98	3,65	

Table 4: Descriptive statistics for numeric variables
Source: Made by the author

Category	Variable	Type	N. Levels	Mode	Occurance percentage per label
Batter	Matchup (B/P)	Categorical	4	R/R	R/R = 37%; L/R = 35%; R/L = 22%; L/L = 6%
	Road/Home	Binary	2	Away	Away = 50%; Home = 50%
	Batter Hand (2)	Binary	2	Right	Right = 59%; Left = 41%
Starting Pitcher	Throwing Hand	Binary	2	Right	Right = 72%; Left = 28%
Ballpark	Roof Type	Categorical	3	Open	Open = 77%; Retractable = 20%; Fixed = 3%
Target Variable	Hit (2)	Binary	2	Yes	Yes = 65%; No = 35%

Table 5: Descriptive statistics for categorical variables
Source: Made by the author

3.2.2.2. Data Visualization

Data visualization is a focal point of any data related word. The concept was popularized by John Tukey (1961) where he defines data analysis as the “procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data”.

In the context of the paper, the best approach was to use both univariate and bivariate visualization techniques. Several types of plots were designed to not only understand the data, but more

importantly to provide a good perception of the facts hidden in the dataset without needing a lot knowledge on the topic of baseball.

Firstly, as seen in table 4, all numeric variables distribution plots were analyzed, these graphs greatly complement other visualization techniques as, for example, it gives a first impression on outliers and their magnitude. Additionally, the distribution plot provides a perception on the distributions of the features, which is essential if further statistical testing is required.

Continuing the topic of outlier detection, boxplots were used to test for outliers in all numeric values. This topic is further explored in the data preparation chapter, where these plots are used in conjunction with some of the descriptive statistics shown previously to detect extreme values that need to be dealt with.

As mentioned in the previous subchapter, data correlation will be of great importance in more than a few processes of the paper. Therefore, in figure 8 are plotted the Pearson's and the Spearman's correlation between the dependent features and the independent features. The correlation values between dependent variables were also calculated, but due to their extensive nature can be found in annexes 8.3 and 8.4.

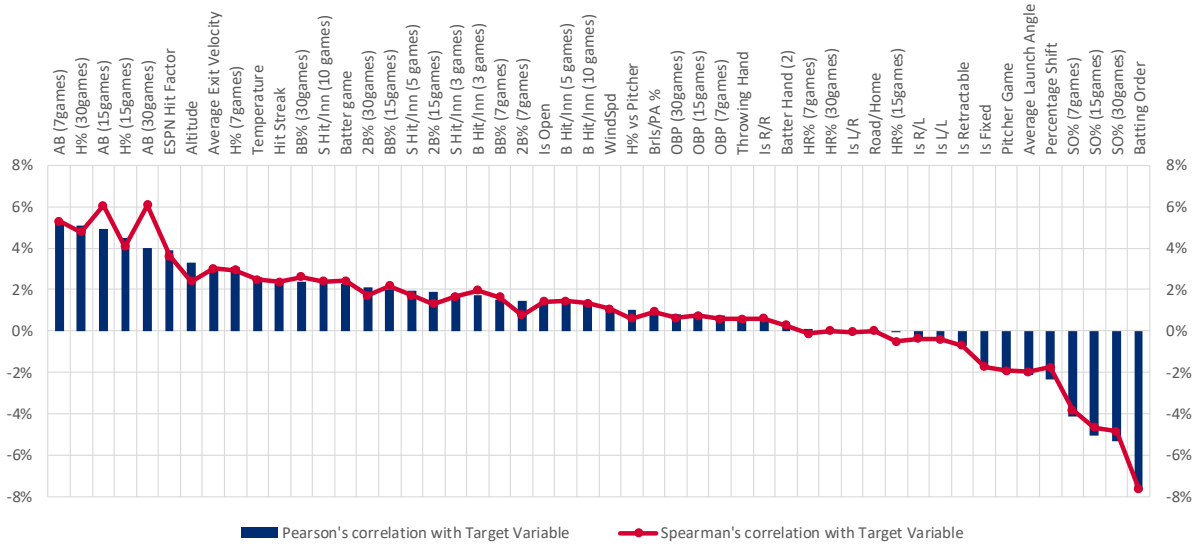


Figure 8: Pearson's and Spearman's correlation for target variable
Source: Made by the author

As seen in the correlation plot, there is no variable with a huge correlation to the target variable but batting order, number of at-bats in previous games and mostly other batter performance variables look to have the most impact on the target variable, at first sight. Nonetheless, these values are very insightful, and it is feasible to select some of top variable, to explore visually.

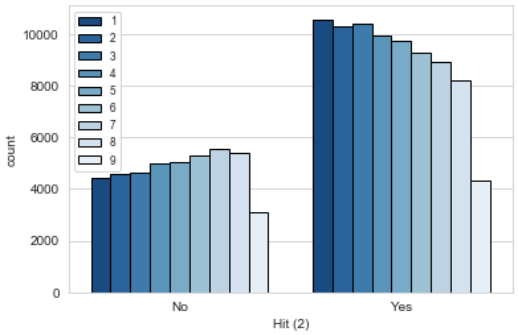


Figure 9: Batting order influence on base hits
Source: Made by the author

Figure 9 depicts the relation between the batting order and if the batter got a base hit in the game. The graph shown matches the negative correlation value of this pair of variables, where the batters in the first positions of the lineup are considerably more likely to get base hit than the bottom of the lineup. This result is a consequence of the fact that the top of lineup bats more often that the bottom of the lineup, thus coaches use these spots for the most talented batters of the team, which usually have the best results. Note that the gap seen in the 9th position is mostly the consequence of removing pitchers batting from the dataset, since very often they bat last in the lineup.

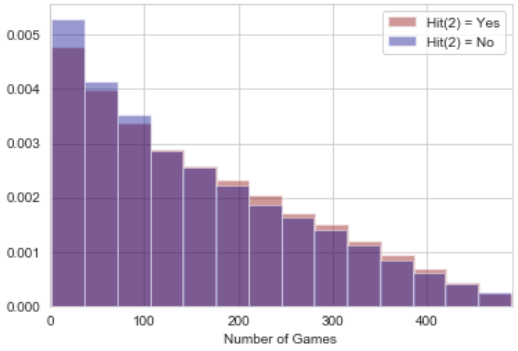


Figure 10: Number of games played by the batter influence on base hits
Source: Made by the author

Another variable taken into consideration when factoring batter performance was the number of games a batter has played. This translates into how experienced a batter is and, as seen in the graph above, there are some insights that are not completely linear. Briefly, a batter with very few games played does not have much success in term of base hits, similarly like batters with a lot of games under their belt. In terms of this feature, there looks to exist a sweet spot or what is commonly known as the prime years of an athlete.

The main factors that affect players of their carrer in baseball are, according to Staszewski & Siegler (1994), physical development which usually peaks at 28-30 years of age, experience which constantly grows during a player's carrer and wear and tear which negatively impacts the player's performance and over the years make a player's more susceptile to injury and to lose of overall performance.

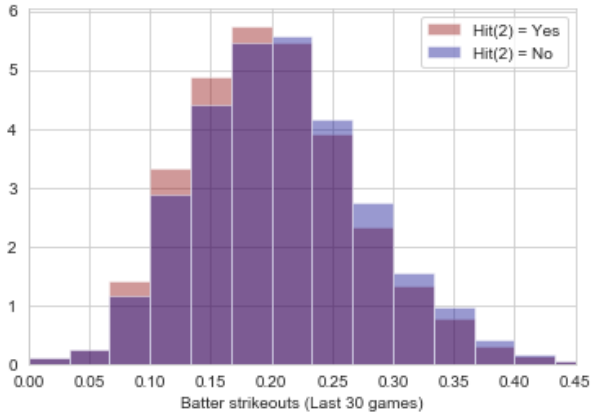


Figure 11: Strikeouts influence on base hits
Source: Made by the author

Figure 11 compares the performance of batters in terms of strikeout percentage in the last 30 games. As expected striking out often translates into less ability to achieve base hits, since this usually means that a batter could not make solid contact with the baseball in his plate appearances. Hence, players which have stricken out in less than 20% of their plate appearances have a greater ability of achieving base hits in a future games.

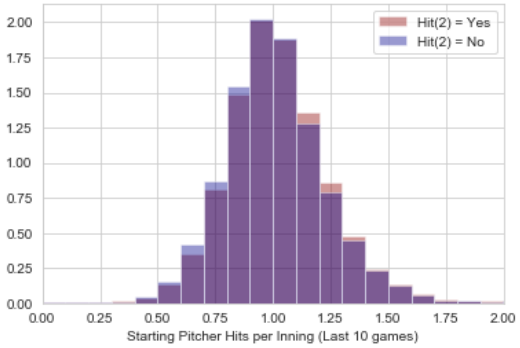


Figure 12: Opponent's starting pitcher performance influence on base hits
Source: Made by the author

Analyzing pitcher's performance, figure 12 illustrates the relationship between the recent performance of starting pitchers and the target variable. The plot shows that starting pitchers which allow approximately more than a base a hit per inning pitched in their games, are more likely to allow base hits in future games. This is a simple measure of performance for starting pitchers and this variable focus on proving insight on what starting pitchers are not performing well.

Finally, figure 13 depicts the impact of the ESPN Hit Factor, related to the different ballparks the MLB games are played in, and the target variable. The plot displays what is expected after seeing the correlation number between the two features in question, but most importantly there are some significant extreme values that should have a deeper analysis.

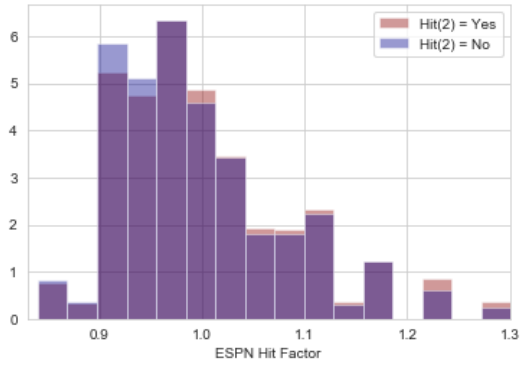


Figure 13: ESPN Hit Factor influence on base hits
Source: Made by the author

The values that lie above the 1.2 in the x-axis refer to the Coors Field, which is a very well-known ballpark for its altitude. This is the home for the Colorado Rockies in Denver, where year after year are achieved above average batting results and Home Runs numbers due to low air density, resulting from a high elevation – 5.200 feet of altitude. In the figure 14 below it is quite visible that there is no ballpark like Coors Field in terms of altitude and ESPN hit factor.

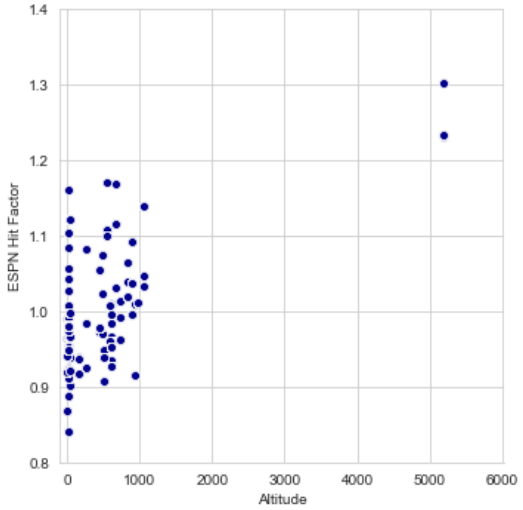


Figure 14: Ballparks displayed by ESPN Hit Factor and Altitude
Source: Made by the author

Additionally, the numbers from the count plot below really show that, not only the ballpark is good for achieving home runs, but it is also very beneficial for base hits. When comparing the games where batters achieved at least a base hit in Coors Field versus the remaining 29 ballparks, there are approximately 7 percentage points of advantage for the former, as seen in figure 15.

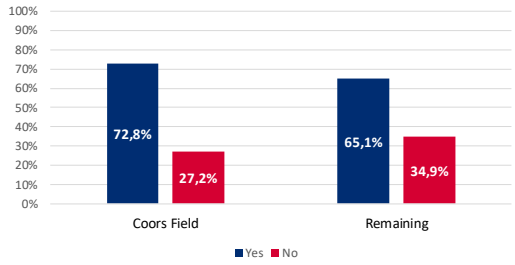


Figure 15: Coors Field versus remaining ballparks, by base hit percentage
Source: Made by the author

Finally, to conclude the analysis of the influence of ESPN Hit Factor on other variables, it was plotted figure 16 where the dataset was grouped by roof type of the ballparks. Using the average values, the conclusion is that open ballparks are the most influenced by the wind and other weather conditions which promote a higher ESPN Hit Factor. This in conjunction with the altitude and the field dimensions are the main aspects that influence the ESPN Hit Factor.

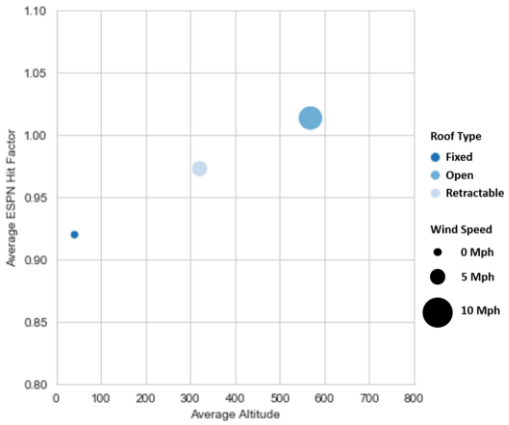


Figure 16: Average windspeed, ESPN Hit Factor and Altitude on ballparks, by type of roof
Source: Made by the author

3.3. DATA PREPARATION

The objective of this chapter is to present the processes performed in the scope of data preparation. Data preparation is crucial for data mining projects, since in the process of collecting and managing data often there are fields that are no longer relevant or are missing and must be dealt with for performance purposes. The prevailing objective of these processes is to minimize garbage in, garbage out (GIGO), i.e. to avoid getting data that is not relevant for the models and would otherwise penalize their performance (Larose, 2005).

3.3.1. Data sampling

Most datasets do not have the same number of observations for each class. The name given to these is imbalanced datasets, for which they do not have a balanced number of observations for each of the classification labels. The prime problem related to imbalanced datasets is false classification metrics, i.e. metrics may skew against or in the direction of the most common label to maximize evaluation metrics (Chawla, 2010).

Having an imbalanced dataset is not an unsurpassable problem, as there are metrics that can go around the problems mentioned previously. Nevertheless, often it is beneficial to transform the dataset, using sampling techniques, into a balanced state. For this purpose, the most common approaches are to either oversample – expand the number of observations from the minority class, or under sample – remove observations from the majority class (Chawla, 2010).

In this paper, the dataset being analyzed is imbalanced, from which the total 155.521 samples, around 65,3% are batters that achieved at least a base hit and the remaining 34,7% are at batters whose game ended without achieving a base hit. Although not very accentuated it possible to determine that the dataset is imbalanced.

Hit	Count	Percentage
Yes	101.619	65,3%
No	53.902	34,7%

Table 6: Distribution of dependent variable
Source: Made by the author

Both the under sample and oversample approaches were taken into consideration for the project. However, oversampling was not a feasible solution in the specific context of this paper, since the objective of the paper is to predict which are the best players for a given day and, therefore the creation of random games with random dates would disturb the analysis. This may have led to players having to play multiple fictitious games in the same days which would not make sense in the context of the regular season of baseball.

In conclusion the only method that will be tested for balancing the dataset is random under sampling. This consists on removing random observations from the majority class until the classes are balanced, i.e. have a similar number of observations for each of the dependent variable values (Chawla, 2010).

3.3.2. Data partitioning

In predictive modelling it is important to understand how to make the most out of the data at hand. In the process of training the dataset we have a two-sided problem, i.e. bias-variance tradeoff. In one hand, a model can have high bias which results in a very broad model that does not capture important relations in the dataset and thus underfits the data. In the other hand, a model may have high variance which results in a strict model that is exceedingly close to the training data and hence overfits the data (Geman & Bienenstock, 1992).

There are a lot of approaches to tackling these problems but one of the most popular is to partition the data. By applying different techniques, we can reduce both variance and bias of the models and thus achieve better overall results.

One of the easiest methods of data partitioning is the holdout method, where we separate the dataset into two subsets - training set and test set. By leaving part of the dataset aside from training we can provide our models with a simulation of the real world and understand their true performance. Nevertheless, taking samples from the training set can lead to a higher bias and thus

one should only use the necessary samples for testing to provide solid evidence of the results achieved (Kohavi, 1995).

Another method commonly used for these purposes is cross-validation, in which the dataset is separated in k number of folds. These folds are used one at the time as the test set as the remaining folds are used for training. Another variation of this method is stratified cross-validation where the process is the same as described previously, but the folds contain approximately the same proportion of depended variable values as the original dataset. The use of these usually leads to a decrease in the variance of the models, i.e. less overfitting (Kohavi, 1995).

Regarding this project both approaches mention above were used. As there was enough data, the initial dataset was firstly divided into training set (80%) and test set (20%) using a simple holdout method. Note that to achieve the best simulation possible, the division was done chronologically, i.e. the first 80% of the games correspond to the training set and the remainder to the test set.

Finally, for the training aspect of the project the training set was recursively divided into a smaller training set (60% of the total) and a validation set (20% of the total). This division implies that for feature selection, hyper parameter tuning, evaluation the data was used with a stratified cross-validation technique with 10-folds. As seen below, figure 17 represents an overview of the partitions and their use for the project.

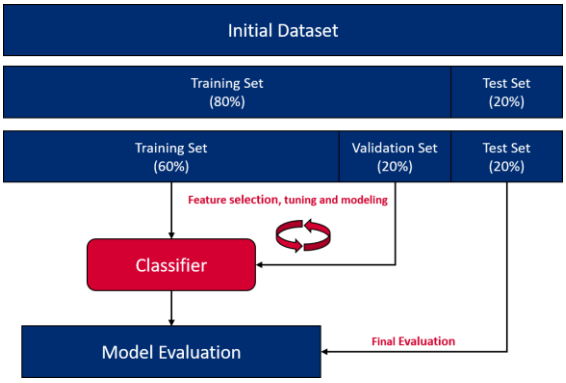


Figure 17: Data partitioning diagram
Source: Made by the author

3.3.3. Data transformation

To build a data mining models the raw variables might not be enough to achieve optimal results. An important aspect of a data mining model is what features are used and how they are treated. The term feature engineering is the process of transforming the raw data into the most useful features possible for the task. It is crucial to have a good understanding of the topic and have experience on working with the types of features in question to apply the best and most correct transformations (Domingos, 2012).

3.3.3.1. Variable transformation

For this project, the variables that were directly extracted from the data source did not held the potential to predict future events, as most of these variables are statistics or events relative to a specific game. The strategy adopted to solve this problem was to link the players performance to the

date dimension. As seen in table 7 below, most performance variables derive from basic statistics that apply player by player, i.e. by representing how well a player has played in the last x games we can achieve a reasonable understanding on future performance.

Category	Variable	Description	Transformation
Batter	Matchup (B/P)	Batter's hand and pitchers hand	Concatenation of Batter Hand and Pitcher Hand
	H% (7games)	Hitting percentage of the batter in the last 7 games played	H (last 7)/PA (last 7)
	H% (15games)	Hitting percentage of the batter in the last 15 games played	H (last 15)/PA (last 15)
	H% (30games)	Hitting percentage of the batter in the last 30 games played	H (last 30)/PA (last 30)
	SO% (7games)	Strikeout percentage of the batter in the last 7 games played	SO (last 7)/PA (last 7)
	SO% (15games)	Strikeout percentage of the batter in the last 15 games played	SO (last 15)/PA (last 15)
	SO% (30games)	Strikeout percentage of the batter in the last 30 games played	SO (last 30)/PA (last 30)
	BB% (7games)	Base on ball percentage of the batter in the last 7 games played	BB (last 7)/PA (last 7)
	BB% (15games)	Base on ball percentage of the batter in the last 15 games played	BB (last 15)/PA (last 15)
	BB% (30games)	Base on ball percentage of the batter in the last 30 games played	BB (last 30)/PA (last 30)
	2B% (7games)	Double percentage of the batter in the last 7 games played	2B (last 7)/PA (last 7)
	2B% (15games)	Double percentage of the batter in the last 15 games played	2B (last 15)/PA (last 15)
	2B% (30games)	Double percentage of the batter in the last 30 games played	2B (last 30)/PA (last 30)
	HR% (7games)	Home run percentage of the batter in the last 7 games played	HR (last 7)/PA (last 7)
	HR% (15games)	Home run percentage of the batter in the last 15 games played	HR (last 15)/PA (last 15)
	HR% (30games)	Home run percentage of the batter in the last 30 games played	HR (last 30)/PA (last 30)
	AB (7games)	Number of at bats in the last of the batter in the last 7 games	Sum of at bats (last 7)
	AB (15games)	Number of at bats in the last of the batter in the last 15 games	Sum of at bats (last 15)
	AB (30games)	Number of at bats in the last of the batter in the last 30 games	Sum of at bats (last 30)
	Hit Streak	Batter's current hitting streak	If last game Hits > 0 then sum +1; Else 0
	H% vs Pitcher	Batter's hit percentage versus this game starting picther	H (all time against pitcher)/ PA (all time against pitcher)
Batting Team	OBP (7games)	Batter's team on base percentage in the last 7 games	H + BB + HBP (last 7)/PA (last 7)
	OBP (15games)	Batter's team on base percentage in the last 15 games	H + BB + HBP (last 15)/PA (last 15)
	OBP (30games)	Batter's team on base percentage in the last 30 games	H + BB + HBP (last 30)/PA (last 30)
Starting Pitcher	Hit/Inn (3 games)	Hit per inning allowed by the opponent's pitcher in the last 3 games	Hits allowed (last 3)/Innings Pitcher (last 3)
	Hit/Inn (5 games)	Hit per inning allowed by the opponent's pitcher in the last 5 games	Hits allowed (last 5)/Innings Pitcher (last 5)
	Hit/Inn (10 games)	Hit per inning allowed by the opponent's pitcher in the last 10 games	Hits allowed (last 10)/Innings Pitcher (last 10)
Bullpen	Hit/Inn (3 games)	Hit per inning allowed by the opponent's bullpen in the last 3 games	Hits allowed (last 3)/Innings Pitcher (last 3)
	Hit/Inn (5 games)	Hit per inning allowed by the opponent's bullpen in the last 5 games	Hits allowed (last 5)/Innings Pitcher (last 5)
	Hit/Inn (10 games)	Hit per inning allowed by the opponent's bullpen in the last 10 games	Hits allowed (last 10)/Innings Pitcher (last 10)

Table 7: Variable transformation description and calculations
Source: Made by the author

3.3.3.2. Min-Max normalization

In data mining there is a global need to normalize numerical variables. This need results from the sensitivity of some algorithms to the range variables possess, thus in cases where this process is not done we can achieve biased results. The most common methods to obtain normalization of the variables is through normalization, standardization and scaling (Larose, 2005).

According to D.T. Larose (2005), the min-max normalization rescales variables in a range between 0 and 1, i.e. the biggest value per variable will assume the form of 1 and the lowest 0. It is considered to be an appropriate method for normalizing datasets where its variables assume different ranges and, at the same time, solves the mentioned problem of biased results for some specific algorithms that cannot manage variables with different ranges.

$$X^* = \frac{X - \min(X)}{range(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Equation 4: Min-Max Normalization Technique
Source: Made by the author, adapted from (Larose, 2005)

3.3.3.3. Binary encoding for categorical variables

Certain algorithms cannot cope with categorical variables. For this reason, if it is intended the use of these types of variables in all models there is a need to transform these into a numerical form. Binary encoding is one of the solutions for this problem, where by creating new columns with binary values it is possible to translate categorical information into 1's and 0's (Larose, 2005).

Note that for categorical variables that do not have an order and assume a lot forms this process is often complex and might be very heavy computationally. For this project encoding was used for only two variables – Roof Type and Matchup, which assume only 3 and 4 different values, respectively. In the end, there are five more columns than before, which is a manageable tradeoff if the variables prove to be useful for the models. Below in table 8, it is possible to see an example of how the encoding works for the variable Roof Type.

Original Roof Type	Fixed	Open	Retractable
Fixed	1	0	0
Open	0	1	0
Retractable	0	0	1

Table 8: Example of variable encoding for the Roof Type Variable
Source: Made by the author

3.3.4. Missing values

Insuring data quality is one of the most important factors when building a data mining model. One of the major concerns to be had when preprocessing data is to understand whether our dataset is plagued by missing values. According to Allison (2001), missing values is data that is missing for some observations but not for all observations of a variable.

The existence of missing values in datasets can affect the diversity of models to be tested, since some algorithms cannot handle this absence of information. Therefore, as part of the preprocessing chapter one should understand how missing values are affecting the dataset and find solutions to counteract their effects (Batista & Monard, 2002).

There are three different types of missing values missing completely at random (1) – occur independently from values of their and other features in the dataset, missing at random (2) – occur independently from values of their feature but are somewhat correlated with another feature in the dataset, missing not at random (3) – they are somewhat correlated to their own feature, i.e. there is a pattern in the feature itself that might explain these missing values (Allison, 2001).

The easiest way to deal with missing values is to ignore them and (1) use models that can handle missing values, such as some decision trees algorithms or (2) ignore the missing values completely and delete the observations which contain these observations. Although the previous solutions might be correct for some cases they have some drawbacks, as limiting the number of algorithms or instances in the dataset in often not a perfect solution.

Therefore, there are other solutions that can help cope with missing values and that overcome some of the drawbacks presented by the previous solutions. Larose (2005), indicates three commonly used methods for replacing the missing values with some other form of value:

1. Replace the missing value with a constant.
2. Replace the missing value with the mean of the feature (for numeric variables) or mode (for categorical variables).
3. Replace the missing value with a value generated at random on par with the feature observed distribution.

Once again there is no correct answer for all missing values problems and studying the features with missing observations is essential to understand which of the methods is the best for the specific case.

Regarding this paper, there is a very limited number of variables with missing values. Originally, only the Statcast variables (Average Launch Angle, Average Exit Velocity, Brls/PA% and Percentage Shift) had missing values, comprising around 1% of the features. These originated from the difference from the two data sources, i.e. some players were not in the Statcast database and therefore did not have a match when building the final database. To solve this issue the observations with missing values relative to Statcast features were deleted due to their immaterial size.

Additionally, some missing values were created after the variable transformation process. These missing values are the calculated performance statistics for the first game of every player in the dataset (pitcher or batter). In sort, every observation which comprises the first game of a player its statistics from the previous "X games" will be NaN since its their first game in the dataset. These occurrences represented around 4% of the dataset for every batter and pitcher.

Although, the number of observations with missing values in these conditions is rather high, to replace these observations it would have been extremely hard to the huge amount of variance in these events and therefore these observations were merely used to build the remaining of the dataset and were not for modelling purposes.

Finally, the variable H% vs pitcher, which calculates the success of a batter versus a specific pitcher suffers more heavily from the previous problem. Around 41% of the observations miss this feature, as it is rarer to have data for these events due to the number of pitchers and batters in the league. It was tested several ways to improve the quality of this variable through replacement of the missing values and in the end the solution chosen was to use the mean of this feature to fill the missing observations.

3.3.5. Outliers

According to Han, Pei & Kamber (2011) an outlier, in the data mining field, is a data object which deviates significantly from the rest of the objects. In other words, outliers are low frequency values that are far away from the remaining data points for a feature or data in general. Additionally, they lie near the limits of the data range, usually in the form of maximum or minimums for a certain feature (Grubbs, 1974, Pyle, 1999, Larose, 2005).

Outliers make a big impact in the process of creating a data mining model, due to their impact in certain algorithms. These algorithms are heavily influenced by extreme values and therefore outliers can introduce a level of bias that is unwanted, resulting in a longer training time, less accuracy and poorer and less robust results. Outliers can be a result of different events, such as machine fault, corruption, human error, natural deviations, etc. However, not extreme values should be labelled as outliers without proper analysis, sometimes is in these values that lie the best information (Larose, 2005).

As seen in the chapter regarding data exploration, there were two main methods to detect outliers in this project. In a first phase, it was calculated the z-score for all features with the objective of understanding what the variables with the most extreme values were considering the mean and standard deviation of the respective features. In this analysis values that are less than -3 standard

deviations or greater than 3 standard deviations than the mean are usually considered outliers. In the context of our problem we cannot simply remove all observations with features that exceed these values, but are a good start on understanding which variables have more outliers and their overall dimension in the context of the whole dataset (Larose, D., & Larose, C., 2014)

$$z = \frac{x - \mu}{\sigma}$$

Equation 5: Z-score formula
Source: Made by the author, adapted from (Larose, D., & Larose, C., 2014)

Another method commonly used for outlier detection is the use of boxplots to visualize feature values in the context of their interquartile ranges. In this method any values that lie under the 1st quartile by more than 1,5 times the size of the interquartile range or over the 3rd quartile by more than 1,5 times the size of the interquartile range is considered an extreme value in the context of the respective feature (McGill, Turkey & Larsen, 1978, Larose, D., & Larose, C., 2014).

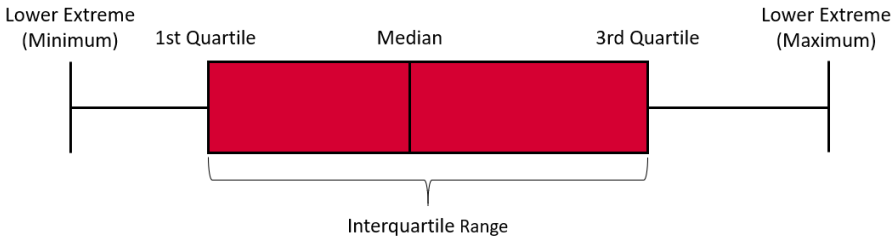


Figure 18: Boxplot example
Source: Made by the author, adapted from (McGill, Turkey & Larsen, 1978)

In the end, the outlier detection of this project consisted on removing the most extreme values from feature using both the information from the z-score analysis and the visualization power of the box-plots. Note that due to the unknown influence of the outliers on the project, every model was tested with outliers and without the identified extreme values. With this it will be possible to have a good comparison of the performance of the models for both scenarios.

3.4. MODELLING

This chapter focus on presenting the methods used for modelling, feature selection and their respective hyper parameter tuning. The objective of this paper is to build a predictive model using classification algorithms, i.e. to classify a new and unknow value of interest given known observations of other relatable variables. In this sense, there are described several processes that will be used to build the final models, with the objective of finding the model that better fits the dataset in question (Larose, 2005, Hand, Mannila & Smyth, 2001)

3.4.1. Algorithms

A. Logistic regression

The logistic regression was developed by David Cox in 1958. This algorithm is an extension of the linear regression, which is mainly used in data mining for the modeling of regression problems. The

former differs from the latter since it looks to solve classification problems by using a sigmoid function or similar to transform the problem into a binary constraint (Cox, 1958).

In SKlearn, most algorithms use a gradient descent approach to the problem, i.e. after the calculation of the probability for an instance the algorithm then proceeds to improve the logistic function base on the error calculated. Doing this for an enough number of iterations/epochs will guarantee a fair conclusion to the problem (Scikit-learn, 2018a).

Finally, the SAGA algorithm was chosen during the hyper parameter tuning for the logistic regression. SAGA follows the path of other algorithms like SAG, also present in the SKlearn library, as an incremental gradient algorithm with fast linear convergence rate. The additional value that SAGA provides is that it supports non-strongly convex problems directly, i.e. without much alterations it better adapts to these types of problems in comparison with other similar algorithms (Defazio, Bach & Lascoste-Julien, 2014)

B. Multi-layer perceptron

According to Shalev-Shwartz & Ben-David (2014), a multi-layer perceptron is a type of neural network, inspired by the structure of the human brain. These algorithms make use of nodes or neurons which connect to one another in different levels with weights attributed to each connection. Additionally, some nodes receive extra information through bias values that are also connected with a certain weight. Overall neural networks are considered quite powerful for classification problems, where its main advantage is its capacity of solving non-linear problems (Zhang, Patuwo, & Hu, 1997, Mitchell, 1997).

The process of training a multi-layer perceptron is solved using an optimization method such as he gradient descent algorithm. By using enough iterations, the forward activation and backpropagation it is possible to iteratively adjust the weights that connect the nodes and achieve a good outcome (Mitchell, 1997).

The multi-layer perceptron used in the SKlearn library uses one hidden layer, according to Palit & Popovic (2005) this strategy can solve most of complex practical problems. The optimization algorithm used for training purposes was Adam, a stochastic gradient-based optimization method which works very well with big quantities of data and provides at the same time low computational drawbacks (Kingma & Ba, 2015). Regarding the activation functions, during the hyper parameter tuning the ones selected were ‘identity’- a no-op activation, which returns f(x) = x and ‘relu’- the rectified linear unit functions, which returns f(x) = max (0, x).

C. Random forest

The concept of random forest is drawn from a collection of decision trees. Decision trees are a simple algorithm with data mining applications, where a tree shaped model progressively grows splitting into branches based on the information held by the variables. Random forests are an ensemble of many of these decision trees, i.e. bootstrapping many decision trees achieves a better overall result, because decision trees are quite prone to overfitting the training data (Kam Ho, 1995).

D. Stochastic gradient descent

The stochastic gradient descent is a common method for optimizing the training of several machine learning algorithms. As mentioned it is used in both the multi-layer perceptron and logistic regression approach available in the SKlearn library. This away it is possible to use the gradient descent as the method of learning, where the loss, i.e. the way the error is calculator during training, is associated with another machine learning algorithm. This enables more control on the optimization and less computational drawbacks for the cost of a higher number of parameters (Scikit-learn, 2018b, Mei, Montanari & Nguyen, 2018).

During the parameter tuning, the loss function that was deemed most efficient was 'log' associated with the logistic regression. Therefore, for this algorithm the error used for training will resemble a normal logistic regression, already described previously.

3.4.2. Feature selection

In data mining projects where there are high quantities of variables it is good practice to reduce the dimensionality of the dataset. Some of the reasons that make this process worthwhile are a reduction in computational processing time and, for some algorithms, overall better results. The latter results from the elimination of the curse of dimensionality – the problem caused by the exponential growth in volume related to adding several dimensions to the Euclidean space (Bellman, 1957). In conclusion, feature selection looks to eliminate variables with reductant information and keeping the ones who are most relevant to the model (Guyon & Elisseeff, 2003).

A. Recursive Feature Elimination

The first method used for selection the optimal set of variables was to use the recursive feature elimination function in SKlearn. In a first instance, all variables are trained, and a coefficient is calculated for each variable, giving the function a value on which features are the best contributors for the model. Thereafter, the worst variable is removed from the set and the process is repeated iteratively until there are no variables left (Guyon, Weston, Barnhill & Vapnik, 2002). Note that the metric chosen for evaluation purposes was area under curve (AUC) and the process was carried out with stratified 10-fold cross validation. The results in figure 19 show the AUC for each number of variables for each of the algorithms capable of applying this function in SKlearn.

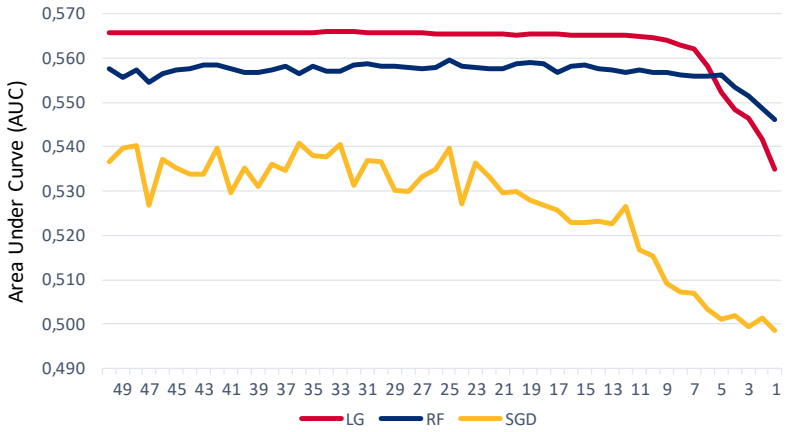


Figure 19: RFE results by algorithm, by number of variables used
Source: Made by the author

B. Principal Component Analysis

Another method used for dimensionality reduction was the principal component analysis (PCA). This technique looks to explain the correlation structure of the features by using a smaller set of linear combinations or components. By combining correlated variables, it is possible to use the predictive power of several variables in a reduced number of components. The main challenge of using the PCA is to choose the correct number of components, this is directly related to the objective of the project as the number of components depends on the amount of variance needed to solve the problem (Larose, D., & Larose, C., 2015, Tipping & Bishop, 1999).

By obtaining the eigenvectors and eigenvalues of the transformed dataset it is possible to calculate the explained variance for each component. According to the Pearson's criterion, 80% of explained variance is a reasonably good approximation for a dataset, accordingly to find the number of needed components to achieve the 80% threshold simply pick the n best components until the summed variance surpasses the defined threshold (Jolliffe, 2002). Figure 20 depicts the four scenarios taken in account in the project, showing the summed explained variance by the number of components.

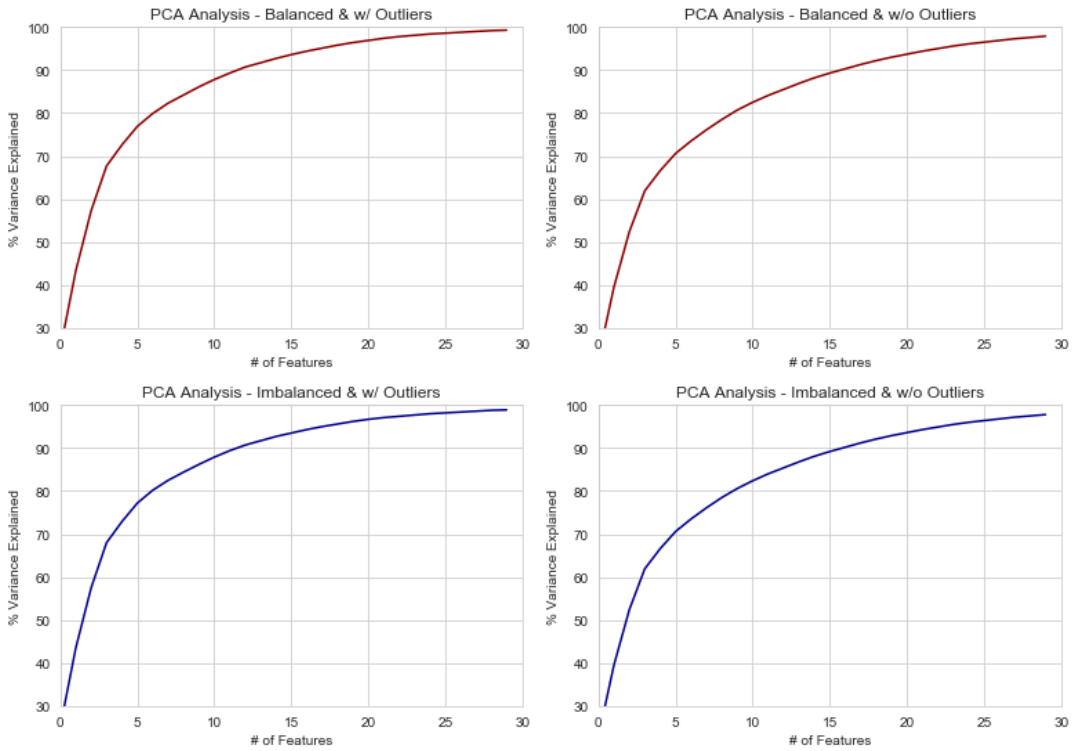


Figure 20: Number of principal components features by summed explained variance
Source: Made by the author

C. Correlation

During the data visualization chapter, correlation between the independent features and the dependent variable were visualized as a mean to understand what the most relevant variables for the models might be. This analysis was carried for all variables, i.e. it was calculated the correlation between all dependent variables as well. This had the objective of doing a correlation-based feature selection, meaning that it is desirable to pick variables highly correlated with the dependent variable and the same time with low intercorrelation with the other independent features (Witten, Frank, & Hall, 2011). This approach is a good method of improving the performance of the models since it takes out variables that do not had value and that could add bias and instability to the model, resulting in inaccurate results (Larose, 2005).

The two main uses of this process were for the variable selection for the multi-layer perceptron variables and for the creation of smaller subsets for all sets previously selected, i.e. removing intercorrelated variables from the set of variables selected by the RFE.

3.4.3. Hyperparameter tuning

A common trait from data mining algorithms is that they are parameterized by a set of hyperparameters. These parameters are used to configure various aspects of the algorithms and, when tuned appropriately, can lead to vastly different models and consequently better results. The main problem of this type of processes is to find a balance between results and computational processing, since it is possible to test every possible combination of hyperparameters but with high computational expenses (Claesen & Moor, 2015).

For this project, the method chosen was Gridsearch with stratified 10-fold cross validation implemented using the SKlearn library. The process is similar to a brute force approach, where Python runs every possible combination of hyperparameters assigned and return as the output the best combination for the predefined metric (Scikit-learn, 2018c). Needing to compare unbalanced and balanced datasets the metric chosen was area under curve. The four tables below illustrate the values inputted for testing and the outputs from the search, for models not using and using PCA.

Parameters	Definition (Sklearn)	Values tested	Best Results	Best Results (PCA)
Solver	Algorithm to use in the optimization problem	Liblinear, Saga, Newton-cg, Lbfgs & Sag	Saga	Saga
C	Inverse of regularization strength.	1 - 200	1	1
Penalty	Used to specify the norm used in the penalization	L1 & L2	L1	L1

Table 9: Hyper parameter tuning for the Logistic Regression
Source: Made by the author

Parameters	Definition (Sklearn)	Values tested	Best Results	Best Results (PCA)
Solver	The solver for weight optimization.	Adam & SGD	Adam	Adam
Alpha	L2 penalty (regularization term) parameter.	0,1 - 0,00000001	0,00001	0,1
No. nodes in hidden layer	The ith element represents the number of neurons in the ith hidden layer.	10 - 20	10	19
Learning rate	Learning rate schedule for weight updates.	Constant & Inscaling	Constant	Invscaling
Activation function	Activation function for the hidden layer.	Identity, Logistic, Tanh & Relu	Identity	Relu

Table 10: Hyper parameter tuning for the Multi-layer Perceptron
Source: Made by the author

Parameters	Definition (Sklearn)	Values tested	Best Results	Best Results (PCA)
Criterion	The function to measure the quality of a split.	Gini & Entropy	Entropy	Entropy
Max depth	The maximum depth of the tree.	5, 10, 15, 20, 25, 30, 35, 40, 45, 50 & 55	10	5
Max features	The number of features to consider when looking for the best split.	Log2, Sqrt & None	Sqrt	Log2
N estimators	The number of trees in the forest.	50, 75 & 100	100	100
Warm start	When set to True, reuse the solution of the previous call to fit and add more estimators to the ensemble, otherwise, just fit a whole	True & False	True	True

Table 11: Hyper parameter tuning for the Random Forest
Source: Made by the author

Parameters	Definition (Sklearn)	Values tested	Best Results	Best Results (PCA)
Loss	The loss function to be used.	Hinge, Log, Modified Huber, Squared Hinge & Perceptron	Log	Log
Alpha	Constant that multiplies the regularization term.	0,1 - 0,0001	0,0001	0,01
Eta0	The initial learning rate for the 'constant', 'invscaling' or 'adaptive' schedules.	0,1, 0,25, 0,5, 0,75 & 1	0,1	1
Llearning rate	The learning rate schedule.	Constant, Optimal & Invscaling	Invcaling	Optimal
Penalty	The penalty (aka regularization term) to be used.	L1, L2 & Elasticnet	Elasticnet	Elasticnet
Power t	The exponent for inverse scaling learning rate.	0,1, 0,25, 0,5, 0,75 & 1	0,25	0,25

Table 12: Hyper parameter tuning for the Steep Gradient Descent
Source: Made by the author

3.5. EVALUATION

The final step of the model process was to choose the metrics that better fitted the problem. The main constraints of the problem were to find good metrics that enabled the comparison between balanced and imbalanced datasets and, of course to achieve the objective of the project. The most appropriate metric to fulfill these requirements was precision with the objective of defining a tight threshold to secure a very high rate of correct predictions on players that would get a base hit. Nevertheless, other metrics that adapt to this type of problem were also calculated in order to get a better overall view of the final models.

A. Area under curve

This metric is related to the Receiver Operating Characteristics (ROC) curve which plots the relationship between sensitivity – the relative frequency of correctly classified positive examples, and specificity – relative frequency of correctly classified negative examples. Overall, the larger the area under the curve (AUC) the best, where an AUC equal to 0.5 is a model that provides no valuable predictions and equal to 1 is a model which predicted every instance correctly (Kononenko & Kukar, 2007).

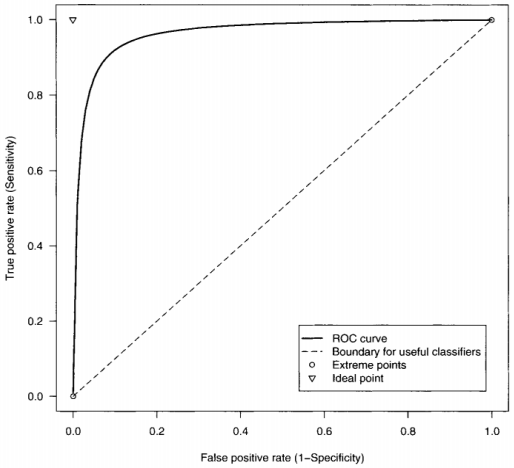


Figure 21: ROC Curve
Source: Retrieved from (Kononenko & Kukar, 2007)

B. Cohen kappa

Cohen's kappa is a metric that expresses the degree of agreement between two annotators on a classification problem. This metric leads to an understanding on how well the model is performing compared to random chance, where if Cohen's kappa equals to 0 the model is probably not performing above the expected random change threshold. (Artstein & Poesio, 2008, Scikit-Learn, 2018d).

$$\kappa = (p_o - p_e)/(1 - p_e)$$

Equation 6: Cohen's Kappa calculation
Source: Retrieved from (Scikit-Learn, 2018d)

C. Precision and average precision

Precision is calculated by diving the amount of actually true instances, or true positives, by the total amount of cases predicted as being of that class, or false positives (Kononenko & Kukar, 2007).

$$Precision = \frac{TP}{TP + FP}$$

Equation 7: Precision calculation
Source: Retrieved from (Kononenko & Kukar, 2007)

Additionally, it was calculated the average precision of the models which comprises the precision per each value of recall. According to scikit-learn (2018e) documentation: “Average Precision summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with increase in recall from the previous threshold used as the weight”.

$$\mathrm{AP} = \sum_n (R_n - R_{n-1}) P_n$$

Equation 8: Average precision calculation
Source: Retrieved from (Scikit-Learn, 2018e)

Finally, for each model it was calculated the precision for the Top 250 and Top 100 instances for each model. This analysis resembles the strategy that will be applied in the real world, for which only the top predictions will be chosen for the game. Note that this analysis will also give a very good impression on what threshold should be used for this point onward.

4. RESULTS AND DISCUSSION

The objective of the paper was to create a model capable of consistently picking MLB players who will get a base hit on a given day. With this in mind, a dataset was built from scratch with variables that, according to the literature review, proved to have the most potential on predicting the aforementioned outcome.

During the process of developing the final 48 models, some alternatives were tested or pondered but later dropped. Starting with sampling, no oversampling technique was used due to not realistically making sense on the context of the work. The main problem with these techniques is that the dataset is time and geographically wise sensitive and, consequently, some unrealistic scenarios would be introduced, for example a pitcher/batter playing at two places at the same time or even play multiple games in same day. This would lead to unreliable or ambiguous conclusions and, at the same time, would be make the day-by-day performing statistics harder to calculate and sometimes misleading.

In terms of other algorithms, Support Vector Machine, K-Nearest Neighbors and Naive Bayes were also considered as possible solutions in the beginning of the project but were deemed unviable. The first two because of high computational run time, often taking days to run some of the processes described in the methodology. The latter did not align with the objective of the problem, since Naive Bayes is known to be a good predictor but a bad estimator, which makes the task of choosing the best predictions harder (Zhang, 2004)

Regarding evaluation metrics, several options were considered, nevertheless, only AUC, Cohen's kappa, average precision, precision were chosen. These metrics were the ones that that the most sense in the context of the problem with special relevance for precision. Combining the probability estimation from the different models, it was possible to simulate a real-life situation and only observe the most probable instances from each model, as it will be done when applying the model. These simulations culminated on the Top 250/100 most probable instances precision metric, which will also help on finding the optimal threshold for the best models selected.

The results can be seen in annexes 8.1 and 8.2, these show some variance when all factors are considered but overall it is viable to retain one main insight from an analysis on the test set metrics. As seen in figure 22, the use of PCA did not help the models perform better according to any metric.

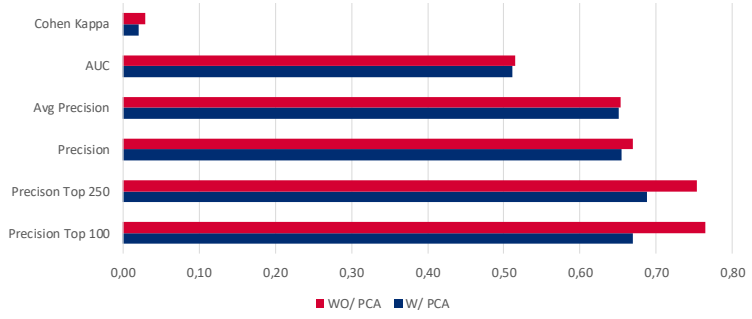
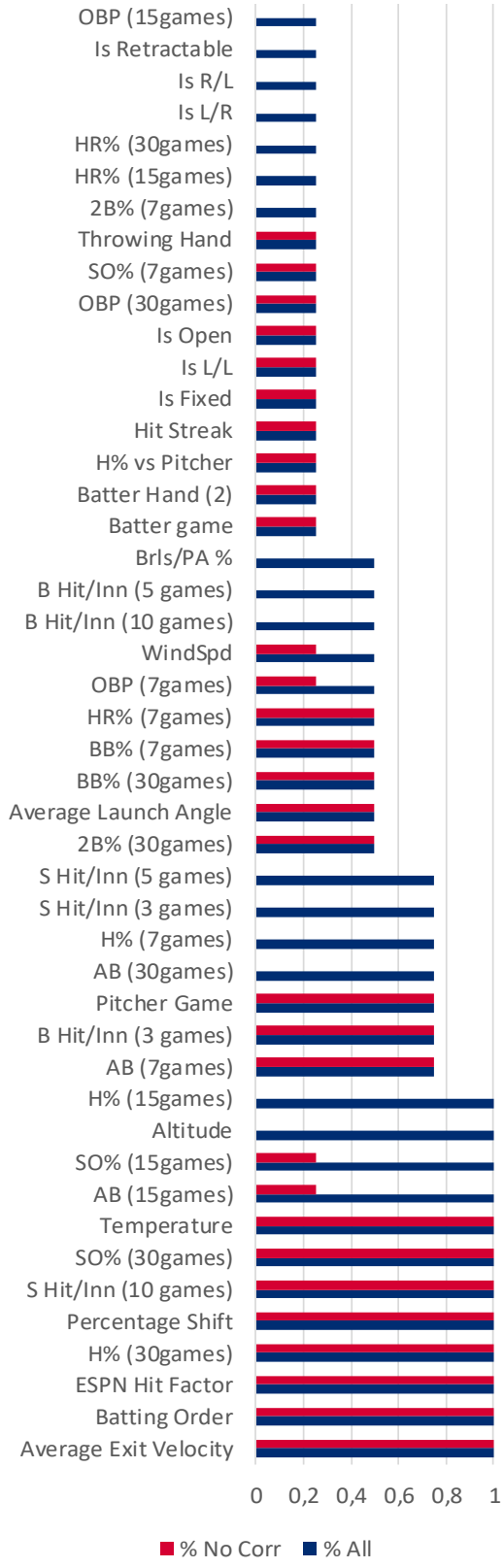


Figure 22: Average model performance on test set, by use of PCA
Source: Made by the author



Apart from PCA, RFE and correlation-based feature selection were also used during the process of selecting the most valuable variables for the models. Figure 23 depicts the most used variables, for the models using all variables selected by RFE and for the subset of variables, which exclude the inter-correlated variables from the RFE selected sets. Note that PCA is not included in this analysis since all variables are used in this process but later decomposed into principle component features, hence it is not logical to make a direct comparison with the other forms of feature selection.

Overall the most used variables fall under the batter performance statistics category, additionally the models use at least one variable from each of the remaining categories, where the most prevalent are hits per innings for the starting pitcher (last 10 games), ESPN hit factor, temperature and hits per innings for the bullpen (last 3 games). The category with the least representation is the team batting statistics, as on-base-percentage (OBP) does not seem to add much prediction value on the outcome of the dependent variable.

This analysis in conjunction with the fact that the smaller subset of variables, i.e. withdrawing the inter-correlated variables, often perform better than the full subset from the original feature selection, highlights some of the variables that are not so relevant when considered with the remaining selected variables. The most predominant variables being cut off, from this process, are batter and pitcher performance statistics and altitude. Since RFE selected several variables that belong to the same subcategory and hence are somewhat correlated. That way it was possible to withdraw some of these excess variables to achieve better overall results, leading to a belief that RFE alone was not the optimal strategy for feature selection for the dataset built for this project.

Figure 23: Variable usage, by type of feature selection
Source: Made by the author

Considering the metrics chosen it was now possible to select the best models, tables 13 and 14 present the various metrics for the top 5 models selected. The logic behind the selection was to choose the models with the highest precision on the Top 100 and 250 instances, giving most attention to those that also performed well on other metrics. The only exception to these rules was the 5th model, which only performed well on the 2 main metrics.

					Validation (10 Strat. Kfold)			
Dataset Balance	**Outliers**	**PCA**	**Variable Selection**	**Model**	**AUC**	**Cohen Kappa**	**Avg Precision**	**Precision**
Random Under Sampling	WO/ Outliers	WO/ PCA	No Variable Correlation	MLP	0,566	0,095	0,555	0,550
Random Under Sampling	WO/ Outliers	WO/ PCA	All Variables	LG	0,567	0,095	0,555	0,548
Random Under Sampling	W/ Outliers	WO/ PCA	No Variable Correlation	SGD	0,562	0,078	0,551	0,539
Random Under Sampling	W/ Outliers	WO/ PCA	All Variables	RF	0,562	0,089	0,551	0,542
Unbalanced Dataset	WO/ Outliers	WO/ PCA	All Variables	LG	0,566	0,000	0,703	0,656

Table 13: Top 5 models evaluation metrics on validation set
Source: Made by the author

					Test set					
Dataset Balance	**Outliers**	**PCA**	**Variable Selection**	**Model**	**AUC**	**Cohen Kappa**	**Avg Precision**	**Precision**	**Precison Top 250**	**Precision Top 100**
Random Under Sampling	WO/ Outliers	WO/ PCA	No Variable Correlation	MLP	0,536	0,057	0,664	0,718	0,760	0,850
Random Under Sampling	WO/ Outliers	WO/ PCA	All Variables	LG	0,528	0,043	0,660	0,716	0,768	0,820
Random Under Sampling	W/ Outliers	WO/ PCA	No Variable Correlation	SGD	0,545	0,080	0,668	0,690	0,760	0,800
Random Under Sampling	W/ Outliers	WO/ PCA	All Variables	RF	0,530	0,061	0,660	0,669	0,776	0,800
Unbalanced Dataset	WO/ Outliers	WO/ PCA	All Variables	LG	0,506	0,015	0,648	0,648	0,784	0,810

Table 14: Top 5 models evaluation metrics on test set
Source: Made by the author

All in all, balanced datasets work well on this project and, most of the top performing models came from sets with random under sampling. When analyzing balanced datasets, these outperform the unbalanced datasets in every metric except for the top 100 and 250 precision, deeming most of them irrelevant. Additionally, imbalanced datasets also had the inconvenience of choosing the most common label for most instances, during validation, producing ambiguous results and limiting possible analysis and conclusions between these results.

Furthermore, methods like outlier's removal, inter-correlated variable removal and the choice of algorithm do not appear to produce dominant strategies in this project and in the right conditions all possibilities produce good results, as seen in the diversity of methods used in the top 5 models.

In a real-life situation, the 1st model in the tables above would provide the best odds of beating the streak with an expected rate of correct picks of 85%, in situations where the model's probability estimate is very high. The remaining top models also prove to be viable, in which models with precision on Top 250 instances higher than the former model give a slight improvement on expected correct picks at lower probability estimations.

The information used to rank the instances by probabilities estimates was further used to calculate at what threshold can we expect to achieve similar correct picks. Note that it is not mandatory to pick a player every single day, therefore it is optimal to wait and pick only when the models are confident enough on when a base hit is occurring. For this analysis figure 24 explores the hit and no hits per probability estimate of each of the top 5 models.

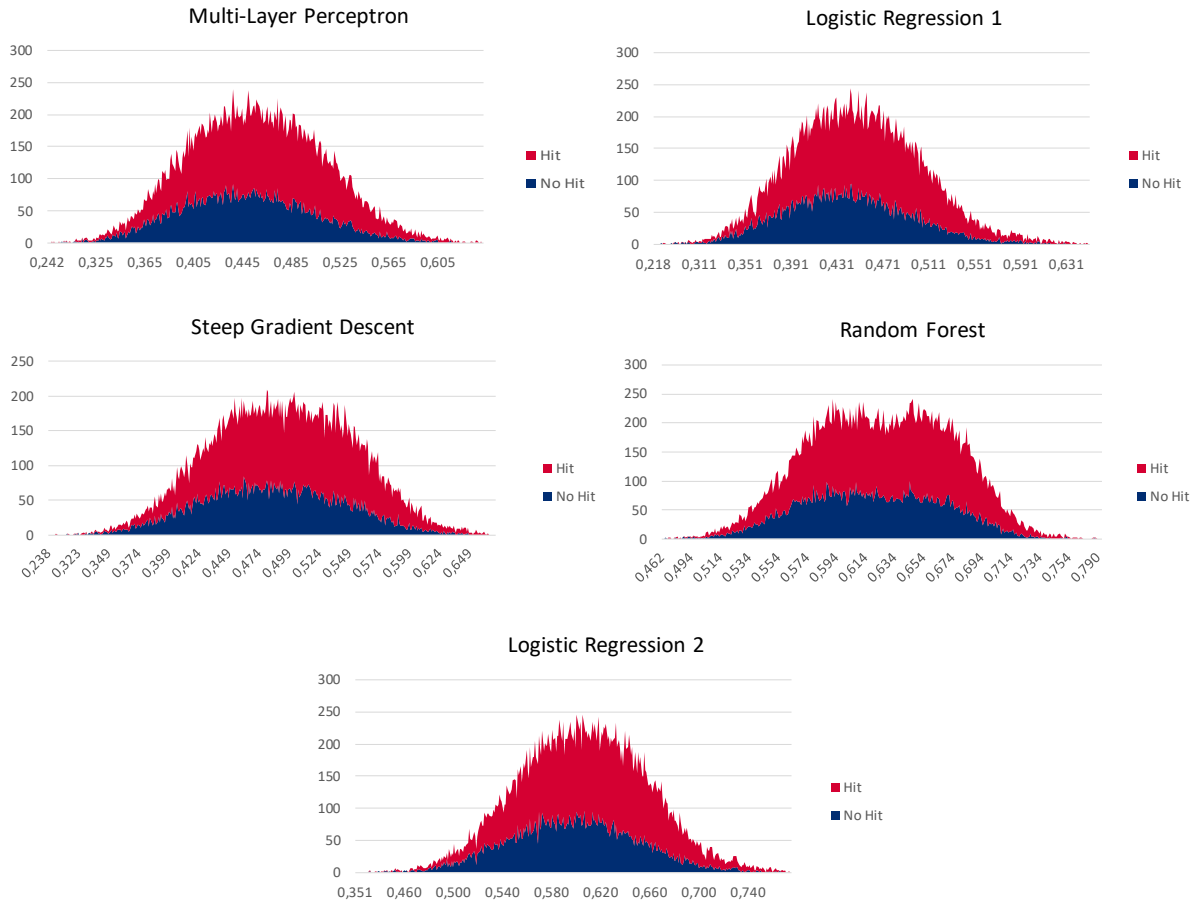


Figure 24: Distribution of probability estimates on top 5 models, by base hit
Source: Made by the author

It possible to note a sweet spot in the far right of each of the figures, where there are areas with no or very few no-hit instances where the models predict hit. After a thorough analysis it is impossible to choose a threshold that gives an 100% change of only picking hit instances that were correctly predicted, for 57 instances. Nevertheless, the top 100 strategy still works well in this analysis and, in table 15, it is depicted the different values that are good thresholds to achieve above 80% expected correct picks.

Probability Estimates	MLP	LG1	SGD	RF	LG2
Maximum probability	0,658	0,671	0,679	0,798	0,792
Minimun probabilty	0,242	0,218	0,238	0,462	0,351
Threshold top 100	**0,608**	**0,616**	**0,643**	**0,743**	**0,749**
Z-score threshold	0,880	0,878	0,918	0,836	0,903
Expected correct ratio	85%	82%	80%	80%	81%

Table 15: Threshold analysis on top 5 models
Source: Made by the author

The models provide different ranges of probabilities but in the end all thresholds fall under approximately 80%-90% of the overall distributions, according to the z-score. With this in mind, some ensembles techniques were tried to improve the expected results, such as majority voting and boosting techniques but none of these techniques provided any improvement in the results and were later dropped.

The final step on the analysis of the results is to finally compare the results from this paper with the papers identified during the literature review:

Paper	Expected Correct Picks	
Random guessing	≈ 60%	
Picking best player	≈ 67%	
Algorithm	**Linear Model**	**MLP**
Goodman & Frey (2013)	70%	-
Clavelli & Gottsegen (2013)	80%	-
Best Models (this paper)	82%	85%

Table 16: Project results versus results of other strategies
Source: Made by the author

The most basic strategies used for playing a game like Beat the Streak is to either pick a complete random player or to pick one of the best batters in the league. These strategies, as expected, have very low results compared to any of the models. From the models identified during the literature review, it is possible to see improvements and strategies that give a player an advantage over the simple strategies. The best and preferred method used by other papers is to use a type of linear model, that was also tried during this project. Nevertheless, the multi-layer perceptron, mentioned in the top 5 algorithms, provides a 5-percentage point improvement over the best model from other papers.

5. CONCLUSIONS

The main objectives of this project were to produce a model that could predict which players were most likely to get a base hit on a given day and, in this sense, provide an estimation of the probability of said event occurring, for the use of stake holders in MLB teams.

To achieve these objectives the following steps were taken:

- Build a database using open-source data including features from a variety of categories;
- Use descriptive statistics and data visualization techniques to explore the value of the features identified during the literature review;
- Build a predictive model using data mining and machine learning techniques, which predicts the probability of a base hit occurring for each instance;
- Apply the model on a test set and analyze the predictions to select the best models and to find the optimal thresholds;

Firstly, the data needed for this project was collected from Baseball Reference, Baseball Savant and ESPN websites. This data was distributed into different Microsoft Excel sheets and later integrated into a single database, displaying the features from each batter's game, not including pitchers batting.

Secondly, the database was imported to Python and structured using the Pandas library. Several descriptive statistics and data visualization techniques were applied to the database, using the Seaborn package to extract insights on the quality of the data, to understand what type of transformations were needed and to gain insights on some the variables being used. Throughout the latter process it was possible to find out that the best variables in terms of correlation to the dependent feature were mostly batting statistics. At the same the variables from this sub-category suffered from inter-correlation with one another, which was taken into consideration during the feature selection. During data visualization some concepts that help batting performance, were confirmed such as lineup position, strikeouts, number of games played and ballpark factors.

Using the insights gained from the second step it was possible to do the preprocessing and transformations on the dataset, making it ready for the third step of the project. Thereafter, several models were built with the main constraints being data set balancing, the use of outliers, the use of PCA and a feature selection using RFE or using RFE in conjunction with a correlation-based feature selection creating a smaller subset of features from the RFE selected variables.

From the final 48 models created it was possible to retain some insights:

- PCA did not perform as well as the other forms of feature selection;
- Overall, balancing the datasets using random under sampling obtained better results than no balancing;
- It was possible to obtain simpler models by removing inter-correlated variables from the RFE selected features and obtain similar or better models;

Finally, after analyzing the performance of the models against the test set, the top 5 models were chosen as possible candidates for usage in a real-world situation. With the ability to calculate the probability estimates for each instance, it was possible to then find the thresholds for each model. The best model gave an expected correct pick rate of 85% on the top 100 picks (precision on top 100 most probable instances), on test set, i.e. half a season worth of instances.

The model that provides these results is a multi-layer perceptron, no outliers, no PCA and with the removal of inter-correlated variables from the original feature selection. When compared to similar works, this model has about 20 percentage points gain on precision over the best basic strategy and a 5 percentage points gain over the best model analyzed during the literature review.

6. LIMITATIONS AND RECCOMENDATIONS FOR FUTURE WORKS

A base hit in baseball has a very close relationship to a player's ability and other factors already mentioned during this paper. Nevertheless, these types of events are prone to be random, since there are a lot of elements that are hard to quantify into features and, thus cannot be fully translated in a machine learning model. The influence of luck can be diminished but it is hard to ever obtain an 100% model in predicting these events. The project at hand had some good results but it is unlikely that with an 85% expected correct pick ratio, it will predict correctly 57 times in a row.

One factor that would improve the performance of the model is the collection of more data. The project was limited in this resource due to using Statcast data that has only been available for the last 4 seasons. Only using 4 season worth of data made it more difficult for the use of variables such specific pitcher vs batter matchups, which is very dependent on a high number of events to be relevant. Furthermore, the Statcast data used in this project was yearly statistics which is lackluster when compared to game-by-game data and, the latter would have made much more sense in the context of this project, but no source was found that offered this option.

In the context of balancing the dataset, it was only experimented the use of random under sampling. The usage of oversampling techniques might prove to be an advantage, but it should be used in a way that does not break the logic behind the regular season of baseball. This was the main limitation of oversampling, since in this project, the problem was tightly related to the time and geographical dimension.

Some variables that were hypothesized in the beginning of the project did not come to fruition. For instance, there were several meteorological variables like humidity, rain and delayed game due to rain, that were not used. These are now available in baseball reference for the last couple of years, but too many instances would have missing values in these features if used, especially for the 2015 and 2016 seasons. If one performs similar models by the 2020 season it is expected to have at least 4 years' worth of information for these variables. Additionally, defensive performance for the opposing team variables were considered as possible valuable features, but in the end, they were quite hard to quantify in any meaningful way and later cut from the project.

In terms of the algorithms used, there is room for improvement in the terms of variety. It would be beneficial to try other solutions like KNN or SVM. These algorithms were developed for a portion of this project but some of the processes like RFE or hyper parameter tuning were too expensive computationally and later were dropped. These algorithms should be run in a simpler manner or in a more powerful machine. Additionally, some simple ensembles were ran using the SKlearn library in Python, but no solution was found to be better than the best models. The usage of ensemble methods is a good opportunity if one is looking to improve its machine learning model and it is very likely that there exists some way to improve this project in that direction.

Finally, it is proposed as future work, the revisiting of this paper and build upon the work presented or even develop a new a predictive model that can overcome the constraints pointed out in this chapter and present a solution with a higher chance of Beating the Streak.

7. REFERENCES

Agyapong, K., Hayfron-Acquah, J., & Asante, M. (2016). An Overview of Data Mining Models (Descriptive and Predictive). *International Journal of Software & Hardware Research in Engineering, 4*, 53-60.

Alamar, B. (2013). *Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers.* Columbia University Press.

Albert, J. (2015). *Improved Component Predictions of Bating.* Department of Mathematics and Statistics Bowling Green State University.

Albert, J. (2016). Improved component predictions of batting and pitching measures. *Journal of Quantitative Analysis in Sports, 12*, 73-85.

Albert, J., & Koning, R. (2008). *Statistical Thinking in Sports.* Chapman & Hall/CRC.

Albert, J., Bartroff, J., Blandford, R., Brooks, D., Derenski, J., Goldstein, L., Hosoi, A., Lorden, G., Nathan, A., & Smith, L. (2018). *Report of the Committee Studying Home Run Rates in Major League Baseball.* Office of the Commissioner of Baseball (BOC).

Allison, P. (2001). Missing Data. *In Quantitative Applications in the Social Sciences*, 72-89.

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics, 34*(4), 555-596.

Bailey, S. (2017). *Forecasting batting averages in MLB.* Burnaby, BC, Canada: MSc master's Project in the Department of Statistics and Actuarial Science, Simon Fraser Universit.

Bailey, S., Loeppky, J., & Swartz, T. (2017). *The Prediction of Batting Averages in Major League Baseball.*

Baseball Reference. (2018). *Baseball Reference*. Retrieved from Baseball Reference: https://www.baseball-reference.com

Baseball Savant. (2018). *Baseball Savant*. Retrieved from Baseball Savant: ttps://baseballsavant.mlb.com

Batista, G., & Monard, M. (2002). Proceedings of the First Internationatal Workshop on Data Cleaning and Preprocessing. *In S. Zhang, Q. Yang & C. Zhang (Eds.), An Analysis of Four Missing Data Treatment Methods for Supervised Learning*, (pp. 142-152).

Beat the Streak (2018). Beat the Streak: Official Rules. Retrieved from Beat the Streak: http://mlb.mlb.com/mlb/fantasy/bts/y2018/?content=rules

Belcastro, L., Marozzo, F., T. D., & Trunfio, P. (2016). Using Scalable Data Mining for Predicting Flight Delays. *ACM Transactions on Intelligent Systems and Technology, 8*(1), 1-20.

Bellman, R. (1957). *Dynamic programming.* Princeton, NJ: Princeton University Press.

Cavanillas, J., Curry, E., & Wahlster, W. (2016). *New Horizons for a Data Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe.* Springer Open.

Chambers, F., Page, B., & Zaidinjs, C. (2003). Atmosphere, weather and baseball: How much farther do baseballs really fly at Denver's Coors Field. *Prof. Geogr. 55th edition*, 491-504.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide.*

Chawla, N. (2010). Data Mining for Imbalanced Datasets: An Overview. In N. Chawla, *Data Mining and Knowledge Discovery Handbook* (pp. 875-886). Springer.

Claesen, M., & Moor, B. (2015). Hyperparameter Search in Machine Learning. *he XI Metaheuristics International Conference*, (pp. 1-5).

Clavelli, J., & Gottsegen, J. (2013). *Maximizing Precision of Hit Predictions in Baseball.*

Collignon, H., & Sultan, N. (2014). *Winning in the Business of Sports.* ATKearney.

Cox, D. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B, 20*(2), 215-242.

Defazio, A., Bach, F., & Lacoste-Julien, S. (2014). SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. *Advances in Neural Information Processing Systems*, *27*, pp. 1-9.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*.

Druschel, H. (2016). *Guide to the Projection Systems.* Retrieved from Beyond the Box Score: https://www.beyondtheboxscore.com/2016/2/22/11079186/projections-marcel-pecota-zips-steamer-explained-guide-math-is-fun

Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. (2002). Modeling and prediction of competitive performance in swimming upon neural networks. *European Journal of Sport Science, 2*(2), 1-10.

ESPN. (2018). *ESPN Hit Factor*. Retrieved from ESPN: http://www.espn.com/mlb/stats/parkfactor

Fast, M. (2010). What the heck is PITCHf/x. *The Hardball Times Annual*, 153-158.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *KDD-96 Proceedings*, (pp. 82-88).

Ffoulkes, P. (2017). *InsideBIGDATA guide to the intelligent use of big data on an industrial scale.* Massachusetts: InsideBIGDATA.

Geman, S., & Bienenstock, E. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation, 4*, 1-58.

Gera, J., Jain, S., Gupta, Sethi, S., D'Silva, S., Munshi, A., Panwala, S., & Pednekar, S. (2016). The Business of Sports, KPMG Article September 2016. Retrieved from https://assets.kpmg.com/content/dam/kpmg/in/pdf/2016/09/the-business-of-sports.pdf

Goodman, I., & Frey, E. (2013). *Beating the Streak: Predicting the MLB Players Most Likely to Get a Hit each Day.*

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research,, 3*, 1157-1182.

Guyon, I., Weston, W., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning, 46*(1-3), 1-3.

Han, J., Pei, P., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (2 ed., Vol. 12). Morgan Kaufmann Publishers, Inc.

Hand, D., Mannila, H. & Smyth, P. (2001). Principles of data mining. Drug safety. *International journal of medical toxicology and drug experience, 30*.

Hauke, J., & Kossowski, T. (2011). Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae, 30*(2), 87-93.

James, B. (2001). *The New Bill James Historical Baseball Abstract.* Free Press.

Jia, R., Wong, C., & Zeng, D. (2013). *Predicting the Major League Baseball Season.*

Jolliffe, I. (2002). *Principal Component Analysi* (2 ed.). Springer.

Kam Ho, T. (1995). Random Decision Trees. Proceedings. *3rd International Conference on Document Analysis and Recognition*, (pp. 278-282).

Kingma, D., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *3rd International Conference for Learning Representations*, (pp. 1-15).

Koch, B., & Panorska, A. (2013). The Impact of Temperature on major League Baseball. *Weather, Climate, and Society journal, 5*(4), 359-366.

Kohavi, R. (1995). Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Articial Intelligence, 5*, 1-7.

Kononenko, I., & Kukar, M. (2007). Measures for Performance Evaluation. In I. Kononenko, & M. Kukar, *Machine Learning and Data Mining: Introduction to Principles and Algorithms* (pp. 66-81). Horwood Publishing Limited.

Kraft, M., & Skeeter, B. (1995). The effect of meteorological conditions on fly ball distances in north American Major League Baseball games. *Geogr. Bull, 37*, 40-48.

Larose, D., & Larose, C. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining* (2 ed.). John Wiley & Sons, Inc.

Larose, D., & Larose, C. (2015). *Data Mining and Predictive Analytics.* John Wiley & Sons, Inc.

Lavalle, S., Lesser, E., S. R., Hopkins, M., & Kruschwitz, N. (2011). Big Data, Analytics and the Path from Insights to Value. *MIT Sloan Management Review, 52*(2), 21-32.

Lewis, M. (2004). *Moneyball: The art of winning an unfair game.* WW Norton & Company.

Mann, R. (2018). *The Marriage of Sports Betting, Analytics and Novice Bettors.* Retrieved from Sports Handle: https://sportshandle.com

Mauboussin, M. (2012). *The Success Equation: Untangling Skill and Luck in Business, Sport, and Investing.* Harvard Business Review Press.

Mcgill, R., T. J., & Larsen, W. (1978). Variations of Box Plots. *The American Statistician, 32*(1), 12-16.

Mei, S., Montanari, A., & Nguyen, P. (2018). A Mean View of the Landscape of Two-Layers Neural Networks. 1-103.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Keribin, C., & Thirion, B. (2011). *A supervised clustering approach for fMRI-based inference of brain states.* Patt Rec.

Mitchel, T. (1997). *Machine learning.* McGraw-Hill Science/Engineering/Math.

MLB. (2018a). *Glossary / Standard Stats / Hit (H)*. Retrieved from MLB: http://m.mlb.com/glossary/standard-stats/hit

MLB. (2018b). *Glossary / Statcast*. Retrieved from MLB: http://m.mlb.com/glossary/statcast

Mordor Intelligence. (2018). *2018Sports Analytics Market - Segmented by End User (Team, Individual), Solution (Social Media Analysis, Business Analysis, Player Fitness Analysis), and Region - Growth, Trends and Forecast (2018 - 2023).*

Morgan, S., Williams, M., & Barnes, C. (2013). Applying decision tree induction for identification of important attributes in one-versus-one player interactions: A hockey exemplar. *Journal of sports sciences, 31*(10), 1031-1037.

Ockerman, S., & Nabity, M. (2014). Predicting the Cy Young Award Winner. *PURE Insights, 3*(1), 9.

Oliver, D. (2004). *Basketball on Paper: Rules and Tools for Performance Analysis.* Potomac Books.

Palit, A., & Popovic, D. (2005). *Neural Networks Approach. InComputational Intelligence in Time Series Forecasting: Theory and Engineering Applications (Advances in Industrial Control)* (1 ed.). London: Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J, Passos, A., & Cournapeua, D. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Pyle, D. (1999). Data Preparation for Data Mining. *Journal on the Theory of Ordered Sets and Its Applications, 17*.

Reep, C., & Benjamin, B. (1968). Skill and chance in association football. *Journal of the Royal Statistical Society. Series A (General), 131*, 581-585.

SAS. (2018). *Machine Learning: What it is and why it matters.* Retrieved from SAS: https://www.sas.com/en_us/insights/analytics/machine-learning.html

Scikit-Learn (2018a). Logistic Regression. Retrieved from: https://scikit-learn.org/stable/modules/linear_model.html#id29

Scikit-Learn (2018b). Stochastic Gradient Descent. Retrieved from: https://scikit-learn.org/stable/modules/sgd.html#classification0

Scikit-Learn (2018c). Grid Search CV. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV

Scikit-Learn (2018d). Average Precision Score. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html

Scikit-Learn (2018e). Cohen's Kappa. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research, 12*(1), 217-222.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms.* Cambridge Univerity Press.

Sheinin, D. (2017). *Astros' World Series win may be remembered as the moment analytics conquered MLB for good.* Retrieved from Washington Post: https://www.washingtonpost.com/sports/astros-world-series-win-may-be-remembered-as-the-moment-analytics-conquered-mlb-for-good/2017/11/02/ac62abaa-bfec-11e7-97d9-bdab5a0ab381_story.html?noredirect=on&utm_term=.e3e732f4790b

Sievert, C., & Mills, B. (2016). Using Publicly Available Baseball Data to Measure and Evaluate Pitching Performance. *Albert/ Handbook of Statistical Methods and Analysis in Sport*, 39-66.

Staszewski, J., & Siegler, R. (1994). The relationship between age and major league baseball performance: Implications for development. *Psychology and Aging, 9*(2), 274-286.

Stekler, H., Sendor, D., & Verlander, R. (2010). Issues in sports forecasting. *nternational Journal of Forecasting, 26*(3), 606-621.

Sykora, M., Chung, P., Folland, J., Halkon, B., & Edirisinghe, E. (2015). Informatics Research Computational Intelligence in Information Systems. In M. Sykora, P. Chung, J. Folland, B. Halkon, & E. Edirisinghe, *Advances in Sports* (pp. 265-274). Springer.

Tipping, M., & Bishop, C. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B, 61*(3), 611-622.

Tukey, J. W. (1961). *The Future of Data Analysis.* Annals of the Institute of Statistical Mathematics.

Valero, C. (2016). Predicting Win-Loss outcomes in MLB regular season games – A comparative study using data mining methods. *International Journal of Computer Science in Sport, 15*(2), 91-112.

Wang, X. (2009). Intelligent Quality Management Using Knowledge Discovery in Databases. *2009 International Conference on Computational Intelligence and Software Engineering*, 1-4.

Witten, I., Frank, E., & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (2 ed.). Morgan Kaufmanne, Inc.

Wolf, G. (2015). The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball by Benjamin Baumer and Andrew Zimbalist (review). *Journal of Sport History, 42*(2), 239-241.

Yuan, L., Liu, A., Yeh, A., Kaufman, A., Reece, A., Bull, P., Franks, A., Wang, S., Illushin, D., & Bornn, L. (2015). A mixture-of-modelers approach to forecasting NCAA tournament outcomes. Journal of Quantitative Analysis in Sports, 11(1), pages 13-27.

Zhang, G., Patuwo, B., & Hu, M. (1997). Forecasting with artificial Neural Networks: The state of state of the art. *International Journal of Forecasting, 14*, 35-62.

Zhang, H. (2004). The Optimality of Naïve Bayes. *Proceeding of the Seventeenth International Florida Artificial Intelligence Research Society Conference.*

8. ANNEXES

8.1. MODELLING EVALUATION METRICS FOR VALIDATION SET

					Validation (10 Strat. Kfold)			
Dataset Balance	Outliers	PCA	Variable Selection	Model	AUC	Cohen Kappa	Avg Precision	Precision
Random Under Sampling	W/ Outliers	W/ PCA	All Variables	LG	0,558	0,084	0,549	0,541
Random Under Sampling	W/ Outliers	W/ PCA	All Variables	MLP	0,559	0,087	0,550	0,541
Random Under Sampling	W/ Outliers	W/ PCA	All Variables	RF	0,556	0,081	0,547	0,539
Random Under Sampling	W/ Outliers	W/ PCA	All Variables	SGD	0,558	0,083	0,549	0,543
Random Under Sampling	W/ Outliers	WO/ PCA	All Variables	LG	0,566	0,095	0,553	0,548
Random Under Sampling	W/ Outliers	WO/ PCA	All Variables	MLP	0,565	0,091	0,554	0,541
Random Under Sampling	W/ Outliers	WO/ PCA	All Variables	RF	0,562	0,089	0,551	0,542
Random Under Sampling	W/ Outliers	WO/ PCA	All Variables	SGD	0,564	0,079	0,551	0,536
Random Under Sampling	W/ Outliers	WO/ PCA	No Variable Correlation	LG	0,565	0,095	0,553	0,548
Random Under Sampling	W/ Outliers	WO/ PCA	No Variable Correlation	MLP	0,565	0,093	0,553	0,550
Random Under Sampling	W/ Outliers	WO/ PCA	No Variable Correlation	RF	0,561	0,089	0,549	0,543
Random Under Sampling	W/ Outliers	WO/ PCA	No Variable Correlation	SGD	0,562	0,078	0,551	0,539
Random Under Sampling	WO/ Outliers	W/ PCA	All Variables	LG	0,560	0,087	0,549	0,544
Random Under Sampling	WO/ Outliers	W/ PCA	All Variables	MLP	0,560	0,088	0,550	0,543
Random Under Sampling	WO/ Outliers	W/ PCA	All Variables	RF	0,558	0,086	0,548	0,542
Random Under Sampling	WO/ Outliers	W/ PCA	All Variables	SGD	0,558	0,085	0,547	0,542
Random Under Sampling	WO/ Outliers	WO/ PCA	All Variables	LG	0,567	0,095	0,555	0,548
Random Under Sampling	WO/ Outliers	WO/ PCA	All Variables	MLP	0,567	0,090	0,555	0,550
Random Under Sampling	WO/ Outliers	WO/ PCA	All Variables	RF	0,559	0,084	0,549	0,541
Random Under Sampling	WO/ Outliers	WO/ PCA	All Variables	SGD	0,564	0,078	0,553	0,543
Random Under Sampling	WO/ Outliers	WO/ PCA	No Variable Correlation	LG	0,567	0,096	0,555	0,548
Random Under Sampling	WO/ Outliers	WO/ PCA	No Variable Correlation	MLP	0,566	0,095	0,555	0,550
Random Under Sampling	WO/ Outliers	WO/ PCA	No Variable Correlation	RF	0,561	0,090	0,549	0,544
Random Under Sampling	WO/ Outliers	WO/ PCA	No Variable Correlation	SGD	0,565	0,081	0,554	0,544
Unbalanced Dataset	W/ Outliers	W/ PCA	All Variables	LG	0,554	0,000	0,692	0,655
Unbalanced Dataset	W/ Outliers	W/ PCA	All Variables	MLP	0,553	0,000	0,692	0,655
Unbalanced Dataset	W/ Outliers	W/ PCA	All Variables	RF	0,553	0,000	0,692	0,655
Unbalanced Dataset	W/ Outliers	W/ PCA	All Variables	SGD	0,553	0,000	0,692	0,655
Unbalanced Dataset	W/ Outliers	WO/ PCA	All Variables	LG	0,565	0,003	0,701	0,656
Unbalanced Dataset	W/ Outliers	WO/ PCA	All Variables	MLP	0,564	0,002	0,701	0,656
Unbalanced Dataset	W/ Outliers	WO/ PCA	All Variables	RF	0,561	0,001	0,699	0,656
Unbalanced Dataset	W/ Outliers	WO/ PCA	All Variables	SGD	0,562	0,002	0,699	0,656
Unbalanced Dataset	W/ Outliers	WO/ PCA	No Variable Correlation	LG	0,565	0,002	0,701	0,656
Unbalanced Dataset	W/ Outliers	WO/ PCA	No Variable Correlation	MLP	0,564	0,002	0,701	0,656
Unbalanced Dataset	W/ Outliers	WO/ PCA	No Variable Correlation	RF	0,561	0,002	0,699	0,656
Unbalanced Dataset	W/ Outliers	WO/ PCA	No Variable Correlation	SGD	0,561	0,001	0,699	0,656
Unbalanced Dataset	WO/ Outliers	W/ PCA	All Variables	LG	0,559	0,002	0,698	0,656
Unbalanced Dataset	WO/ Outliers	W/ PCA	All Variables	MLP	0,560	0,000	0,699	0,656
Unbalanced Dataset	WO/ Outliers	W/ PCA	All Variables	RF	0,557	0,000	0,696	0,656
Unbalanced Dataset	WO/ Outliers	W/ PCA	All Variables	SGD	0,558	0,000	0,696	0,656
Unbalanced Dataset	WO/ Outliers	WO/ PCA	All Variables	LG	0,566	0,000	0,703	0,656
Unbalanced Dataset	WO/ Outliers	WO/ PCA	All Variables	MLP	0,566	0,002	0,703	0,656
Unbalanced Dataset	WO/ Outliers	WO/ PCA	All Variables	RF	0,559	0,000	0,698	0,656
Unbalanced Dataset	WO/ Outliers	WO/ PCA	All Variables	SGD	0,564	0,001	0,702	0,656
Unbalanced Dataset	WO/ Outliers	WO/ PCA	No Variable Correlation	LG	0,566	0,002	0,702	0,656
Unbalanced Dataset	WO/ Outliers	WO/ PCA	No Variable Correlation	MLP	0,566	0,002	0,702	0,656
Unbalanced Dataset	WO/ Outliers	WO/ PCA	No Variable Correlation	RF	0,559	0,002	0,697	0,656
Unbalanced Dataset	WO/ Outliers	WO/ PCA	No Variable Correlation	SGD	0,564	0,001	0,702	0,656

8.2. Modelling Evaluation Metrics for Test Set

					Test set					
Dataset Balance	Outliers	PCA	Variable Selection	Model	AUC	Cohen Kappa	Avg Precision	Precision	Precison Top 250	Precision Top 100
Random Under Sampling	W/ Outliers	W/ PCA	All Variables	LG	0,526	0,048	0,658	0,669	0,748	0,750
Random Under Sampling	W/ Outliers	W/ PCA	All Variables	MLP	0,524	0,046	0,657	0,666	0,756	0,760
Random Under Sampling	W/ Outliers	W/ PCA	All Variables	RF	0,534	0,063	0,662	0,674	0,748	0,760
Random Under Sampling	W/ Outliers	W/ PCA	All Variables	SGD	0,527	0,052	0,658	0,667	0,756	0,760
Random Under Sampling	W/ Outliers	WO/ PCA	All Variables	LG	0,520	0,048	0,655	0,656	0,724	0,730
Random Under Sampling	W/ Outliers	WO/ PCA	All Variables	MLP	0,544	0,085	0,667	0,679	0,776	0,750
Random Under Sampling	W/ Outliers	WO/ PCA	All Variables	RF	0,530	0,061	0,660	0,669	0,776	0,800
Random Under Sampling	W/ Outliers	WO/ PCA	All Variables	SGD	0,542	0,079	0,666	0,680	0,784	0,750
Random Under Sampling	W/ Outliers	WO/ PCA	No Variable Correlation	LG	0,522	0,053	0,656	0,657	0,732	0,750
Random Under Sampling	W/ Outliers	WO/ PCA	No Variable Correlation	MLP	0,531	0,070	0,660	0,663	0,744	0,690
Random Under Sampling	W/ Outliers	WO/ PCA	No Variable Correlation	RF	0,535	0,062	0,663	0,682	0,704	0,660
Random Under Sampling	W/ Outliers	WO/ PCA	No Variable Correlation	SGD	0,545	0,080	0,668	0,690	0,760	0,800
Random Under Sampling	WO/ Outliers	W/ PCA	All Variables	LG	0,520	0,036	0,655	0,663	0,648	0,650
Random Under Sampling	WO/ Outliers	W/ PCA	All Variables	MLP	0,516	0,031	0,653	0,659	0,568	0,560
Random Under Sampling	WO/ Outliers	W/ PCA	All Variables	RF	0,500	0,000	0,646	0,646	0,592	0,540
Random Under Sampling	WO/ Outliers	W/ PCA	All Variables	SGD	0,530	0,056	0,660	0,671	0,744	0,700
Random Under Sampling	WO/ Outliers	WO/ PCA	All Variables	LG	0,528	0,043	0,660	0,716	0,768	0,820
Random Under Sampling	WO/ Outliers	WO/ PCA	All Variables	MLP	0,533	0,051	0,662	0,720	0,804	0,740
Random Under Sampling	WO/ Outliers	WO/ PCA	All Variables	RF	0,501	0,002	0,646	0,646	0,780	0,790
Random Under Sampling	WO/ Outliers	WO/ PCA	All Variables	SGD	0,506	0,009	0,649	0,735	0,744	0,790
Random Under Sampling	WO/ Outliers	WO/ PCA	No Variable Correlation	LG	0,528	0,043	0,660	0,719	0,744	0,800
Random Under Sampling	WO/ Outliers	WO/ PCA	No Variable Correlation	MLP	0,536	0,057	0,664	0,718	0,760	0,850
Random Under Sampling	WO/ Outliers	WO/ PCA	No Variable Correlation	RF	0,522	0,034	0,656	0,700	0,724	0,760
Random Under Sampling	WO/ Outliers	WO/ PCA	No Variable Correlation	SGD	0,507	0,010	0,649	0,736	0,720	0,790
Unbalanced Dataset	W/ Outliers	W/ PCA	All Variables	LG	0,500	0,000	0,646	0,646	0,740	0,710
Unbalanced Dataset	W/ Outliers	W/ PCA	All Variables	MLP	0,500	0,001	0,646	0,646	0,744	0,750
Unbalanced Dataset	W/ Outliers	W/ PCA	All Variables	RF	0,500	0,000	0,646	0,646	0,748	0,730
Unbalanced Dataset	W/ Outliers	W/ PCA	All Variables	SGD	0,500	0,000	0,646	0,646	0,748	0,690
Unbalanced Dataset	W/ Outliers	WO/ PCA	All Variables	LG	0,500	0,000	0,646	0,646	0,720	0,680
Unbalanced Dataset	W/ Outliers	WO/ PCA	All Variables	MLP	0,501	0,002	0,646	0,646	0,772	0,740
Unbalanced Dataset	W/ Outliers	WO/ PCA	All Variables	RF	0,501	0,003	0,646	0,646	0,768	0,780
Unbalanced Dataset	W/ Outliers	WO/ PCA	All Variables	SGD	0,502	0,006	0,647	0,647	0,788	0,750
Unbalanced Dataset	W/ Outliers	WO/ PCA	No Variable Correlation	LG	0,500	0,000	0,646	0,646	0,724	0,660
Unbalanced Dataset	W/ Outliers	WO/ PCA	No Variable Correlation	MLP	0,500	0,000	0,646	0,646	0,736	0,720
Unbalanced Dataset	W/ Outliers	WO/ PCA	No Variable Correlation	RF	0,500	0,001	0,646	0,646	0,700	0,700
Unbalanced Dataset	W/ Outliers	WO/ PCA	No Variable Correlation	SGD	0,505	0,012	0,648	0,648	0,756	0,800
Unbalanced Dataset	WO/ Outliers	W/ PCA	All Variables	LG	0,500	0,000	0,646	0,646	0,648	0,620
Unbalanced Dataset	WO/ Outliers	W/ PCA	All Variables	MLP	0,500	0,000	0,646	0,646	0,624	0,560
Unbalanced Dataset	WO/ Outliers	W/ PCA	All Variables	RF	0,500	0,000	0,646	0,646	0,524	0,520
Unbalanced Dataset	WO/ Outliers	W/ PCA	All Variables	SGD	0,500	0,000	0,646	0,646	0,680	0,660
Unbalanced Dataset	WO/ Outliers	WO/ PCA	All Variables	LG	0,506	0,015	0,648	0,648	0,784	0,810
Unbalanced Dataset	WO/ Outliers	WO/ PCA	All Variables	MLP	0,504	0,010	0,647	0,647	0,788	0,760
Unbalanced Dataset	WO/ Outliers	WO/ PCA	All Variables	RF	0,501	0,002	0,646	0,646	0,776	0,800
Unbalanced Dataset	WO/ Outliers	WO/ PCA	All Variables	SGD	0,511	0,028	0,651	0,651	0,744	0,820
Unbalanced Dataset	WO/ Outliers	WO/ PCA	No Variable Correlation	LG	0,507	0,018	0,649	0,649	0,756	0,800
Unbalanced Dataset	WO/ Outliers	WO/ PCA	No Variable Correlation	MLP	0,504	0,010	0,647	0,647	0,756	0,810
Unbalanced Dataset	WO/ Outliers	WO/ PCA	No Variable Correlation	RF	0,503	0,007	0,647	0,647	0,760	0,790
Unbalanced Dataset	WO/ Outliers	WO/ PCA	No Variable Correlation	SGD	0,513	0,032	0,652	0,652	0,724	0,800

8.3. FULL PEARSON'S CORRELATION TABLE

Variables	Batter Hand (2)	Throwing Hand	Road/Home	HRvsPitcher	Batter game	Batting Order	HR (7 games)	HR (15games)	HR (30games)	SO% (7games)	SO% (15games)	SO% (30games)	BB% (7games)	BB% (15games)	BB% (30games)	2B% (7games)	2B% (15games)	2B% (30games)	HRR% (7games)
Batter Hand (2)		-0.25	0.00	0.01	-0.02	0.00	0.02	0.03	0.03	-0.04	0.05	0.06	-0.03	-0.04	-0.05	0.00	0.00	0.00	0.03
Throwing Hand	-0.25		0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.02	0.01	0.00	0.00	0.00	0.00	0.00
Road/Home	0.00	0.01		0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
HRvs Pitcher	0.01	0.00	0.00		0.01	-0.04	0.03	0.05	0.07	-0.04	-0.04	-0.05	0.01	0.02	0.03	0.01	0.02	0.02	0.00
Batter game	-0.02	0.00	0.00	0.01		-0.22	0.03	0.04	0.06	-0.10	-0.12	-0.13	0.05	0.08	0.10	0.02	0.02	0.03	0.09
Batting Order	0.09	0.00	0.00	0.00	-0.22		-0.16	-0.22	-0.27	0.16	0.19	0.21	-0.12	-0.17	-0.22	-0.06	-0.09	-0.11	-0.10
HR (7games)	0.02	0.01	0.00	0.03	0.03	-0.16		0.72	0.57	-0.32	-0.34	-0.30	0.44	0.33	0.25	0.40	0.29	0.23	0.27
HR (15games)	0.03	0.01	0.00	0.05	0.04	-0.22	0.72		0.78	-0.27	-0.34	-0.30	0.31	0.46	0.35	0.48	0.40	0.32	0.37
HR (30games)	0.04	0.01	0.00	0.07	0.06	-0.27	0.57	0.78		-0.32	-0.37	-0.33	0.31	0.46	0.35	0.48	0.40	0.32	0.37
SO% (7games)	0.05	-0.02	0.00	-0.04	-0.12	0.19	-0.32	-0.27	-0.32		0.82	0.74	-0.13	-0.18	-0.14	-0.10	-0.13	-0.10	-0.11
SO% (15games)	0.06	-0.02	0.00	-0.04	-0.13	0.21	-0.32	-0.34	-0.37	0.82		0.90	-0.10	-0.14	-0.19	-0.08	-0.10	-0.13	-0.14
SO% (30games)	0.06	-0.02	0.00	-0.05	-0.12	0.19	-0.25	-0.34	-0.33	0.74	0.90		-0.10	-0.14	-0.19	-0.08	-0.10	-0.13	-0.14
BB% (7games)	-0.03	0.01	0.00	0.01	0.05	-0.12	0.44	0.31	0.24	-0.17	-0.13	-0.10		0.68	0.54	0.18	0.13	0.19	0.16
BB% (15games)	-0.04	0.02	0.00	0.02	0.08	-0.17	0.33	0.46	0.35	-0.15	-0.18	-0.14	0.68		0.77	0.13	0.19	0.12	0.12
BB% (30games)	-0.06	0.02	0.00	0.03	0.10	-0.22	0.25	0.35	0.45	-0.13	-0.16	-0.19	0.54	0.77		0.10	0.14	0.19	0.10
2B% (7games)	0.00	0.01	0.00	0.02	0.03	-0.09	0.40	0.48	0.40	-0.10	-0.13	-0.10	0.13	0.19	0.14		0.77	0.77	0.09
2B% (15games)	0.00	0.01	0.00	0.02	0.05	-0.11	0.23	0.32	0.40	-0.10	-0.13	-0.14	0.11	0.16	0.19	0.35		0.77	0.09
2B% (30games)	0.00	0.01	0.00	0.02	0.05	-0.12	0.15	0.22	0.14	0.09	0.11	0.12	0.09	0.12	0.13	0.03	0.09		0.06
HRR% (7games)	0.03	0.00	0.00	0.00	0.09	-0.10	0.27	0.17	0.11	0.04	0.09	0.12	0.15	0.12	0.10	0.03	0.02	0.09	
HRR% (15games)	0.04	0.00	0.00	0.00	0.12	-0.12	0.15	0.22	0.14	0.11	0.12	0.16	0.11	0.16	0.13	0.13	0.03	0.09	0.75
HRR% (30games)	0.04	0.00	0.00	0.00	0.15	-0.14	0.09	0.12	0.16	0.15	0.18	0.20	0.09	0.13	0.16	0.03	0.03	0.09	0.62
AB (7games)	0.00	0.01	0.00	-0.01	0.19	-0.40	0.15	0.26	0.28	-0.09	-0.14	-0.17	-0.01	0.00	0.03	0.10	0.10	0.11	0.06
AB (15games)	0.00	0.01	0.00	0.03	0.21	-0.40	0.12	0.22	0.24	-0.07	-0.12	-0.17	-0.01	0.00	0.03	0.10	0.10	0.11	0.06
AB (30games)	0.00	0.01	0.00	0.03	0.21	-0.38	0.09	0.17	0.24	-0.07	-0.11	-0.17	-0.01	0.00	0.03	0.09	0.10	0.11	0.06
Hit Streak	0.00	0.01	0.00	0.01	0.03	-0.14	0.04	0.06	0.07	-0.03	-0.04	-0.05	0.03	0.03	0.04	0.02	0.03	0.03	0.02
Average Exit Velocity	0.02	-0.01	0.00	-0.02	0.16	0.00	-0.08	-0.12	-0.15	0.13	0.15	0.17	0.01	0.01	0.01	0.05	0.06	0.14	0.08
SLG/PA%	0.06	0.00	0.00	-0.01	0.03	-0.13	0.00	0.00	-0.01	0.22	0.26	0.29	0.06	0.09	0.11	0.04	0.05	0.09	0.32
Percentage Shift	-0.47	0.11	0.00	-0.04	0.16	-0.08	-0.09	-0.13	-0.17	0.13	0.16	0.17	0.02	0.03	0.04	0.00	0.00	0.00	0.20
OBP (7games)	-0.01	0.00	0.00	0.01	0.05	0.00	0.00	0.00	0.01	-0.12	-0.08	-0.06	0.11	0.10	0.08	0.03	0.04	0.05	0.08
OBP (15games)	-0.01	0.00	0.00	0.01	0.05	0.00	0.02	0.04	0.06	-0.10	-0.11	-0.13	0.11	0.13	0.11	0.05	0.06	0.07	0.08
OBP (30games)	-0.01	0.01	0.00	0.01	0.04	0.00	0.02	0.04	0.06	-0.10	-0.11	-0.13	0.11	0.13	0.11	0.05	0.06	0.07	0.08
Pitcher Game	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S: Hit/Ann (3 games)	0.00	0.01	0.01	-0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
S: Hit/Ann (5 games)	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01
S: Hit/Ann (10 games)	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.01
Batter Hand (2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Batter Hand (30games)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Temperature	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
WindSpd	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Altitude	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ESPN Hit Factor	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LF Feet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CF Feet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RF Feet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LF Reach	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CF Reach	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RF Reach	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Line Favorable	-0.03	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Line Unfavorable	-0.03	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Lt-L/R	-0.30	-0.40	0.00	0.00	0.05	-0.02	0.02	0.02	0.02	-0.02	-0.02	-0.02	0.02	0.01	0.04	0.00	0.00	0.01	0.02
Lt-R/R	-0.88	0.46	0.00	-0.01	-0.01	-0.08	-0.03	-0.04	-0.04	-0.04	-0.04	-0.05	0.01	0.02	0.04	0.00	0.00	0.01	0.02
Lt-L/L	0.44	-0.85	0.00	0.00	0.02	0.01	0.02	0.03	0.02	0.00	0.01	0.03	-0.02	-0.01	-0.05	-0.01	-0.01	-0.01	-0.01
Lt-R/L	0.64	0.48	0.00	0.01	0.00	0.08	0.04	0.05	0.05	0.02	0.03	0.03	0.00	-0.01	-0.02	0.01	0.01	0.01	0.04
Hit (2)	0.00	0.01	0.00	0.01	0.02	-0.08	0.03	0.04	0.05	-0.04	-0.05	-0.05	0.01	0.02	0.02	0.01	0.02	0.02	0.00

Variables	2B% (30games)	HRR% (7games)
Batter Hand (2)	0.00	0.03
Throwing Hand	0.00	0.00
Road/Home	0.00	0.00
HRvs Pitcher	0.02	0.00
Batter game	0.05	0.02
Batting Order	-0.11	-0.10
HR (7games)	0.23	0.27
HR (15games)	0.32	0.37
HR (30games)	0.37	0.37
SO% (7games)	-0.11	-0.10
SO% (15games)	-0.14	-0.13
SO% (30games)	-0.13	-0.14
BB% (7games)	0.16	0.15
BB% (15games)	0.19	0.12
BB% (30games)	0.19	0.10
2B% (7games)	0.77	0.09
2B% (15games)	0.77	0.09
2B% (30games)	0.77	0.06
HRR% (7games)	0.09	
HRR% (15games)	0.09	0.75
HRR% (30games)	0.09	0.62
AB (7games)	0.10	0.06
AB (15games)	0.10	0.06
AB (30games)	0.09	0.06
Hit Streak	0.03	0.03
Average Exit Velocity	0.14	0.13
SLG/PA%	0.08	0.23
Percentage Shift	0.12	0.28
OBP (7games)	0.01	0.20
OBP (15games)	0.07	0.32
OBP (30games)	0.07	0.08
Pitcher Game	0.00	0.00
S: Hit/Ann (3 games)	0.00	0.01
S: Hit/Ann (5 games)	0.00	0.01
S: Hit/Ann (10 games)	0.00	0.01
Batter Hand (2)	0.00	0.00
Batter Hand (30games)	0.00	0.00
Temperature	0.00	0.02
WindSpd	0.00	-0.01
Altitude	0.00	0.00
ESPN Hit Factor	0.00	0.00
LF Feet	0.00	0.00
CF Feet	0.00	0.00
RF Feet	0.00	0.00
LF Reach	0.00	0.00
CF Reach	0.00	0.00
RF Reach	0.00	0.00
Line Favorable	0.00	0.00
Line Unfavorable	0.00	0.00
Lt-L/R	0.01	0.02
Lt-R/R	0.00	-0.04
Lt-L/L	-0.01	-0.01
Lt-R/L	-0.01	-0.04
Hit (2)	0.02	0.04

Variables	HR% (15games)	HR% (30games)	AB (7games)	AB (15games)	AB (30games)	Hit Streak	Average Launch Angle	Average Exit Velocity	Brls/PA %	Percentage Shift	OBP (7games)	OBP (15games)	OBP (30games)	Pitcher Game	S Hit/Inn (3 games)	S Hit/Inn (5 games)	S Hit/Inn (10 games)	B Hit/Inn (3 games)	B Hit/Inn (5 games)	B Hit/Inn (10 games)
batter Hand (2)	0,04	0,04	0,00	0,00	0,00	0,00	0,02	0,07	0,06	-0,47	-0,01	-0,01	-0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Throwing Hand	0,00	0,00	0,01	0,02	0,01	0,01	-0,01	0,01	0,00	0,11	0,00	0,01	0,01	-0,02	0,01	0,01	0,01	0,00	0,00	0,01
Road/ Home	0,00	0,00	-0,01	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	-0,04	-0,02	0,01
H% vs Pitcher	0,00	0,00	0,03	0,03	0,02	0,01	-0,02	0,00	-0,01	-0,04	0,01	0,01	0,01	-0,01	0,03	0,04	0,04	0,00	0,01	0,01
batter game	0,12	0,15	0,19	0,31	0,42	0,03	0,16	0,16	0,03	0,16	0,05	0,07	0,10	0,37	-0,01	-0,01	-0,01	0,00	0,00	0,01
Batting Order	-0,12	-0,14	-0,40	-0,36	-0,31	-0,14	0,00	-0,21	-0,13	-0,08	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00
H% (7games)	0,15	0,09	0,25	0,14	0,09	0,48	-0,08	0,05	0,00	-0,09	0,22	0,17	0,13	-0,03	0,01	0,01	0,00	0,05	0,04	0,04
H% (15games)	0,22	0,12	0,26	0,22	0,13	0,37	-0,12	0,06	0,00	-0,13	0,14	0,20	0,17	-0,03	0,00	0,00	0,00	0,04	0,03	0,03
H% (30games)	0,14	0,16	0,28	0,24	0,18	0,29	-0,15	0,07	-0,01	-0,17	0,11	0,15	0,18	-0,03	0,00	0,00	0,00	0,03	0,03	0,02
SO% (7games)	0,11	0,15	-0,09	-0,10	-0,10	-0,17	0,13	0,13	0,22	0,13	-0,12	-0,10	-0,08	0,05	0,00	0,00	0,00	-0,02	-0,02	-0,02
SO% (15games)	0,12	0,18	-0,14	-0,13	-0,12	-0,14	0,15	0,16	0,26	0,16	-0,08	-0,11	-0,09	0,06	0,00	0,00	0,01	-0,02	-0,01	-0,01
SO% (30games)	0,16	0,20	-0,17	-0,17	-0,14	-0,12	0,17	0,18	0,29	0,17	-0,06	-0,08	-0,09	0,06	0,00	0,00	0,01	-0,02	-0,01	-0,01
BB% (7games)	0,11	0,09	-0,01	-0,01	0,01	0,23	0,01	0,09	0,06	0,02	0,11	0,11	0,08	0,00	0,00	0,00	0,01	0,03	0,02	0,02
BB% (15games)	0,16	0,13	0,00	0,00	0,03	0,18	0,01	0,13	0,09	0,03	0,10	0,14	0,13	0,00	0,00	0,00	0,01	0,02	0,02	0,02
BB% (30games)	0,13	0,16	0,03	0,03	0,04	0,15	0,01	0,16	0,11	0,04	0,08	0,11	0,15	-0,01	0,00	0,00	0,01	0,02	0,01	0,02
2B% (7games)	-0,02	-0,01	0,10	0,05	0,03	0,19	0,05	0,07	0,04	0,00	0,11	0,09	0,07	-0,01	0,00	0,01	0,00	0,02	0,02	0,02
2B% (15games)	-0,01	0,00	0,10	0,08	0,04	0,14	0,06	0,10	0,05	0,00	0,08	0,11	0,10	0,00	0,00	0,00	0,00	0,02	0,01	0,01
2B% (30games)	0,01	0,01	0,11	0,09	0,07	0,11	0,08	0,12	0,07	0,01	0,07	0,09	0,12	0,00	0,00	0,00	0,00	0,02	0,01	0,01
HR% (7games)	0,75	0,62	0,06	0,05	0,06	0,13	0,23	0,28	0,32	0,20	0,08	0,08	0,06	0,02	0,01	0,01	0,01	0,01	0,01	0,01
HR% (15games)	[redacted]	0,82	0,06	0,07	0,08	0,09	0,30	0,37	0,42	0,26	0,06	0,09	0,08	0,03	0,01	0,01	0,01	0,01	0,01	0,01
HR% (30games)	0,82	[redacted]	0,06	0,08	0,10	0,06	0,37	0,45	0,51	0,32	0,05	0,07	0,09	0,04	0,01	0,01	0,01	0,01	0,01	0,01
AB (7games)	0,06	0,06	[redacted]	0,74	0,52	0,19	-0,03	0,09	0,01	-0,03	0,13	0,11	0,09	0,01	0,00	0,00	0,00	0,04	0,03	0,03
AB (15games)	0,07	0,08	0,74	[redacted]	0,84	0,13	0,00	0,11	0,01	0,00	0,07	0,11	0,11	0,07	0,01	0,00	0,00	0,02	0,02	0,03
AB (30games)	0,08	0,10	0,52	0,84	[redacted]	0,09	0,04	0,12	0,01	0,04	0,05	0,08	0,11	0,14	0,01	0,01	0,00	0,02	0,02	0,03
Hit Streak	0,09	0,06	0,19	0,13	0,09	[redacted]	-0,03	0,04	0,01	-0,03	0,15	0,11	0,08	-0,02	0,00	0,00	0,00	0,05	0,04	0,03
Average Launch Angle	0,30	0,37	-0,03	0,00	0,04	-0,03	[redacted]	0,23	0,35	0,43	0,01	0,01	0,02	0,07	0,01	0,01	0,01	0,00	0,00	0,00
Average Exit Velocity	0,37	0,45	0,09	0,11	0,12	0,04	0,23	[redacted]	0,64	0,33	0,03	0,03	0,04	-0,01	0,00	-0,01	-0,01	0,00	0,00	0,00
Brls/PA %	0,42	0,51	0,01	0,01	0,01	0,01	0,35	0,64	[redacted]	0,39	0,01	0,02	0,01	-0,07	0,00	0,00	0,00	0,00	0,00	0,00
Percentage Shift	0,26	0,32	-0,03	0,00	0,04	-0,03	0,43	0,33	0,39	[redacted]	0,00	0,01	0,01	0,02	0,00	0,00	0,00	0,00	0,00	0,00
OBP (7games)	0,06	0,05	0,13	0,07	0,05	0,15	0,01	0,03	0,01	0,00	[redacted]	0,70	0,54	0,02	0,01	0,01	0,02	0,15	0,13	0,11
OBP (15games)	0,09	0,07	0,11	0,11	0,08	0,11	0,01	0,03	0,02	0,01	0,70	[redacted]	0,76	0,03	0,02	0,02	0,02	0,11	0,09	0,09
OBP (30games)	0,08	0,09	0,09	0,13	0,11	0,08	0,02	0,04	0,01	0,01	0,54	0,76	[redacted]	0,05	0,02	0,02	0,03	0,08	0,07	0,07
Pitcher Game	0,03	0,04	0,01	0,07	0,14	-0,02	0,07	-0,01	-0,07	0,02	0,02	0,03	0,05	[redacted]	-0,08	-0,09	-0,11	-0,02	-0,02	-0,02
S Hit/Inn (3 games)	0,01	0,01	0,00	0,01	0,01	0,00	0,01	0,00	0,00	0,00	0,01	0,02	0,02	-0,08	[redacted]	[redacted]	0,82	0,68	0,03	0,04	0,06
S Hit/Inn (5 games)	0,01	0,01	0,00	0,00	0,01	0,00	0,01	-0,01	0,00	0,00	0,01	0,02	0,02	-0,09	0,82	[redacted]	0,83	0,03	0,05	0,07
S Hit/Inn (10 games)	0,01	0,01	0,00	0,00	0,00	0,00	0,01	-0,01	0,00	0,00	0,02	0,02	0,03	-0,11	0,68	0,83	[redacted]	0,03	0,05	0,07
B Hit/Inn (3 games)	0,01	0,01	0,04	0,02	0,02	0,05	0,00	0,00	0,00	0,00	0,15	0,11	0,08	-0,02	0,03	0,03	0,03	[redacted]	[redacted]	0,54
B Hit/Inn (5 games)	0,01	0,01	0,03	0,02	0,02	0,04	0,00	0,00	0,00	0,00	0,13	0,09	0,07	-0,02	0,04	0,05	0,05	0,77	[redacted]	0,71
B Hit/Inn (10 games)	0,01	0,01	0,03	0,03	0,03	0,03	0,00	0,00	0,00	0,00	0,11	0,09	0,07	-0,02	0,06	0,07	0,07	0,54	0,71	[redacted]
Temperature	0,02	0,02	0,03	0,07	0,09	0,02	-0,02	0,00	0,00	-0,01	0,03	0,04	0,04	0,03	0,01	0,01	0,00	0,03	0,04	0,05
WindSpd	-0,02	-0,02	0,00	0,00	-0,01	0,00	0,02	-0,02	-0,02	0,01	-0,01	0,00	-0,01	0,00	0,01	0,02	0,02	0,00	0,00	0,01
Altitude	-0,01	-0,01	0,02	0,02	0,02	0,02	-0,03	-0,01	-0,02	-0,03	0,02	0,02	0,03	-0,03	0,01	0,02	0,02	0,05	0,04	0,04
ESPN Hit Factor	0,00	0,00	0,02	0,02	0,01	0,02	-0,02	0,00	-0,01	-0,01	0,04	0,03	0,05	0,02	0,02	0,02	0,02	0,06	0,05	0,03
is Fixed	0,01	0,01	-0,01	-0,01	-0,01	-0,01	0,01	-0,01	0,00	0,01	0,01	0,00	-0,01	-0,01	-0,01	-0,02	-0,03	0,00	0,00	0,00
is Open	-0,02	-0,02	0,01	0,00	0,00	0,01	0,03	-0,02	-0,03	0,01	0,01	0,01	0,00	-0,01	0,02	0,03	0,03	0,02	0,02	0,03
is Retractable	0,01	0,02	0,00	0,01	0,01	-0,01	-0,04	0,03	0,03	-0,01	-0,02	-0,01	0,00	0,01	-0,02	-0,02	-0,02	-0,02	-0,02	-0,03
is L/L	0,02	0,02	0,03	0,03	0,03	0,02	0,00	0,02	0,02	0,17	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
is L/R	-0,05	-0,05	-0,02	-0,01	-0,01	-0,01	-0,02	-0,08	-0,07	0,40	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,01
is R/L	-0,01	-0,01	-0,03	-0,04	-0,03	-0,03	0,01	-0,01	-0,01	-0,22	-0,01	-0,01	-0,02	0,02	-0,01	-0,01	0,00	0,00	0,00	0,00
is R/R	0,04	0,05	0,03	0,03	0,02	0,02	0,02	0,09	0,07	-0,29	0,00	0,00	0,00	-0,02	0,01	0,00	0,00	0,00	0,00	0,00
Hit (2)	0,00	0,00	0,05	0,05	0,04	0,02	-0,02	0,03	0,01	-0,02	0,01	0,01	0,01	-0,02	0,02	0,02	0,02	0,02	0,01	0,01

Variables	Temperature	WindSpd	Altitude	ESPN Hit Factor	Is Fixed	Is Open	Is Retractable	Is L/L	Is L/R	Is R/L	Is R/R	Hit (2)
Batter Hand (2)	0,00	-0,02	0,00	-0,01	0,01	-0,03	0,03	-0,30	-0,88	0,44	0,64	0,00
Throwing Hand	0,01	0,00	-0,01	-0,01	0,00	0,00	-0,01	-0,40	0,46	-0,85	0,48	0,01
Road/Home	0,00	0,00	0,00	0,00	-0,01	0,01	0,00	0,00	0,00	-0,01	0,01	0,00
H% vs Pitcher	0,00	0,00	0,01	0,00	-0,01	0,00	0,00	0,00	-0,01	0,00	0,01	0,01
Batter game	0,05	-0,02	0,02	0,00	-0,01	0,00	0,01	0,05	-0,01	-0,02	0,00	0,02
Batting Order	-0,01	0,00	-0,02	0,00	0,02	-0,01	0,00	-0,02	-0,08	0,01	0,08	-0,08
H% (7games)	0,03	0,00	0,02	0,02	-0,01	0,01	-0,01	0,02	-0,03	-0,02	0,04	0,03
H% (15games)	0,04	0,00	0,03	0,02	-0,01	0,01	-0,01	0,02	-0,04	-0,02	0,05	0,04
H% (30games)	0,04	-0,01	0,04	0,03	-0,02	0,01	0,00	0,02	-0,04	-0,02	0,05	0,05
SO% (7games)	-0,01	-0,01	0,00	0,00	0,02	-0,02	0,01	-0,02	-0,04	0,03	0,02	-0,04
SO% (15games)	-0,01	-0,01	0,00	0,00	0,03	-0,02	0,01	-0,02	-0,04	0,03	0,03	-0,05
SO% (30games)	-0,01	-0,01	0,00	0,00	0,04	-0,02	0,01	-0,02	-0,05	0,03	0,03	-0,05
BB% (7games)	0,00	-0,01	0,01	-0,01	0,00	-0,01	0,01	0,02	0,01	-0,03	0,00	0,01
BB% (15games)	0,00	-0,01	0,01	-0,02	0,00	-0,01	0,01	0,03	0,02	-0,04	-0,01	0,02
BB% (30games)	0,00	-0,01	0,01	-0,03	-0,01	-0,01	0,02	0,04	0,04	-0,05	-0,02	0,02
2B% (7games)	0,00	0,00	0,01	0,01	0,00	0,01	-0,01	0,00	0,00	-0,01	0,01	0,01
2B% (15games)	0,00	0,00	0,02	0,02	0,00	0,00	0,00	0,01	0,00	-0,01	0,01	0,02
2B% (30games)	0,00	0,00	0,02	0,02	-0,01	0,01	0,00	0,01	-0,01	-0,01	0,01	0,02
HR% (7games)	0,02	-0,01	-0,01	0,00	0,01	-0,01	0,01	0,02	-0,04	-0,01	0,04	0,00
HR% (15games)	0,02	-0,02	-0,01	0,00	0,01	-0,02	0,01	0,02	-0,05	-0,01	0,04	0,00
HR% (30games)	0,02	-0,02	-0,01	0,00	0,01	-0,02	0,02	0,02	-0,05	-0,01	0,05	0,00
AB (7games)	0,03	0,00	0,02	0,02	-0,01	0,01	0,00	0,03	-0,02	-0,03	0,03	0,05
AB (15games)	0,07	0,00	0,02	0,02	-0,01	0,00	0,01	0,03	-0,01	-0,04	0,03	0,05
AB (30games)	0,09	-0,01	0,02	0,01	-0,01	0,00	0,01	0,03	-0,01	-0,03	0,02	0,04
Hit Streak	0,02	0,00	0,02	0,02	-0,01	0,01	-0,01	0,02	-0,01	-0,03	0,02	0,02
Average Launch Angle	-0,02	0,02	-0,03	-0,02	0,01	0,03	-0,04	0,00	-0,02	0,01	0,02	-0,02
Average Exit Velocity	0,00	-0,02	-0,01	0,00	-0,01	-0,02	0,03	0,02	-0,08	-0,01	0,09	0,03
Brls/PA %	0,00	-0,02	-0,02	-0,01	0,00	-0,03	0,03	0,02	-0,07	-0,01	0,07	0,01
Percentage Shift	-0,01	0,01	-0,03	-0,01	0,01	0,01	-0,01	0,17	0,40	-0,22	-0,29	-0,02
OBP (7games)	0,03	-0,01	0,02	0,04	0,01	0,01	-0,02	0,00	0,00	-0,01	0,00	0,01
OBP (15games)	0,04	0,00	0,02	0,03	0,00	0,01	-0,01	0,00	0,01	-0,01	0,00	0,01
OBP (30games)	0,04	-0,01	0,03	0,05	-0,01	0,00	0,00	0,00	0,01	-0,02	0,00	0,01
Pitcher Game	0,03	0,00	-0,03	0,02	-0,01	-0,01	0,01	0,00	0,00	0,02	-0,02	-0,02
S Hit/Inn (3 games)	0,01	0,01	0,01	0,02	-0,01	0,02	-0,02	0,00	0,00	-0,01	0,01	0,02
S Hit/Inn (5 games)	0,01	0,02	0,02	0,02	-0,02	0,03	-0,02	0,00	0,00	-0,01	0,00	0,02
S Hit/Inn (10 games)	0,00	0,02	0,02	0,02	-0,03	0,03	-0,02	0,00	0,00	0,00	0,00	0,02
B Hit/Inn (3 games)	0,03	0,00	0,05	0,06	0,00	0,02	-0,02	0,00	0,00	0,00	0,00	0,02
B Hit/Inn (5 games)	0,04	0,00	0,04	0,05	0,00	0,02	-0,02	0,00	0,01	0,00	0,00	0,01
B Hit/Inn (10 games)	0,05	0,01	0,04	0,03	0,00	0,03	-0,03	0,00	0,01	0,00	0,00	0,01
Temperature		-0,11	0,08	0,11	-0,03	0,02	0,00	0,00	0,01	-0,01	0,00	0,03
WindSpd	-0,11		-0,01	0,15	-0,28	0,52	-0,43	0,00	0,01	0,00	-0,02	0,01
Altitude	0,08	-0,01		0,60	-0,09	0,14	-0,10	0,01	0,00	0,00	-0,01	0,03
ESPN Hit Factor	0,11	0,15	0,60		-0,19	0,27	-0,20	0,02	0,00	0,01	-0,01	0,04
Is Fixed	-0,03	-0,28	-0,09	-0,19		-0,34	-0,09	0,00	-0,01	0,00	0,01	-0,02
Is Open	0,02	0,52	0,14	0,27	-0,34		-0,90	0,00	0,03	-0,01	-0,02	0,01
Is Retractable	0,00	-0,43	-0,10	-0,20	-0,09	-0,90		0,00	-0,03	0,01	0,02	-0,01
Is L/L	0,00	0,00	0,01	0,02	0,00	0,00	0,00		-0,18	-0,13	-0,19	0,00
Is L/R	0,01	0,01	0,00	0,00	-0,01	0,03	-0,03	-0,18		-0,39	-0,56	0,00
Is R/L	-0,01	0,00	0,00	0,01	0,00	-0,01	0,01	-0,13	-0,39		-0,41	0,00
Is R/R	0,00	-0,02	-0,01	-0,01	0,01	-0,02	0,02	-0,19	-0,56	-0,41		0,01
Hit (2)	0,03	0,01	0,03	0,04	-0,02	0,01	-0,01	0,00	0,00	0,00	0,01	

8.4. FULL SPEARMAN'S CORRELATION TABLE

Variable	Batter Hand (2)	Throwing Hand	Road/Home	HRs vs Pitcher	Batter game	Batting Order	HRs (7games)	HRs (15games)	HRs (30games)	50% (7games)	50% (15games)	50% (30games)	BB% (7games)	BB% (15games)	BB% (30games)	28% (7games)	28% (15games)	28% (30games)	HR% (7games)
Batter Hand (2)		-0.24	0.00	0.01	-0.02	0.09	0.02	0.03	0.04	0.04	0.05	0.06	-0.03	-0.04	-0.06	0.01	0.01	0.00	0.03
Throwing Hand	-0.24		0.01	0.00	0.00	0.00	0.01	0.01	0.01	-0.01	-0.02	-0.02	0.01	0.02	0.02	0.00	0.00	0.00	0.00
Road/Home	0.00	0.01		0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01	-0.03	0.01	0.00	0.00	0.00	0.00	0.00
HRs vs Pitcher	0.01	0.00	0.01		0.02	0.00	0.03	0.05	0.07	0.02	0.04	0.05	0.01	0.02	0.01	0.01	0.01	0.01	0.04
Batter game	0.09	0.00	0.01	0.02		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Batting Order	-0.08	0.00	0.00	0.00	0.00		0.03	0.05	0.07	0.02	0.04	0.05	0.01	0.02	0.01	0.01	0.01	0.01	0.04
HRs (7games)	0.03	0.01	0.00	0.03	0.00	0.03		0.71	0.54	0.02	0.04	0.05	-0.03	-0.04	-0.06	0.01	0.01	0.00	0.03
HRs (15games)	0.03	0.01	0.00	0.05	0.00	0.05	0.71		0.76	0.03	0.05	0.07	-0.04	-0.05	-0.07	0.01	0.01	0.00	0.04
HRs (30games)	0.04	0.01	0.00	0.07	0.00	0.07	0.54	0.76		0.04	0.07	0.09	-0.05	-0.07	-0.09	0.01	0.01	0.00	0.05
50% (7games)	0.04	-0.01	0.00	-0.03	0.00	0.16	-0.30	-0.26	-0.36		0.81	0.72	-0.16	-0.14	-0.12	-0.10	-0.09	-0.09	-0.09
50% (15games)	0.05	-0.02	0.01	-0.04	0.00	0.22	-0.26	-0.31	-0.36	0.81		0.89	-0.13	-0.17	-0.15	-0.07	-0.07	-0.11	-0.12
50% (30games)	0.06	-0.02	0.01	-0.05	0.00	0.22	-0.26	-0.31	-0.36	0.72	0.89		-0.08	-0.11	-0.14	-0.05	-0.05	-0.09	-0.11
BB% (7games)	-0.03	0.01	-0.03	0.01	0.00	0.01	0.01	0.01	0.01	0.02	0.01	-0.08		0.66	0.46	0.11	0.15	0.11	0.17
BB% (15games)	-0.04	0.02	-0.01	0.02	0.00	0.01	0.01	0.01	0.02	0.02	0.01	-0.11	0.66		0.75	0.17	0.15	0.12	0.19
BB% (30games)	-0.06	0.02	-0.01	0.02	0.00	0.01	0.01	0.01	0.02	0.02	0.01	-0.13	0.46	0.75		0.17	0.15	0.12	0.19
28% (7games)	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.11	0.17	0.17		0.68	0.74	0.00
28% (15games)	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.15	0.15	0.15	0.68		0.74	0.01
28% (30games)	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.17	0.17	0.17	0.74	0.74		0.01
HR% (7games)	0.03	0.00	0.00	0.03	0.00	0.03	0.04	0.04	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	
HR% (15games)	0.04	0.00	0.00	0.04	0.00	0.04	0.05	0.05	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
HR% (30games)	0.05	0.00	0.00	0.05	0.00	0.05	0.06	0.06	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
AB (7games)	0.00	0.01	0.00	-0.01	0.00	-0.01	0.02	0.02	0.03	-0.01	-0.01	-0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
AB (15games)	0.00	0.02	0.00	0.00	0.00	-0.01	0.02	0.02	0.03	-0.01	-0.01	-0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
AB (30games)	0.01	0.01	0.00	0.00	0.00	-0.01	0.02	0.02	0.03	-0.01	-0.01	-0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Hit Streak	0.00	-0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg Exit Velocity	0.02	-0.01	0.00	-0.02	0.00	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg Launch Angle	0.02	0.00	0.00	-0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Barrels/PA %	0.05	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Percentage Shift	-0.00	0.11	0.00	-0.03	0.18	-0.08	0.07	0.08	0.10	0.15	0.15	0.21	0.03	0.04	0.05	0.01	0.02	0.02	0.01
OBP (7games)	-0.01	0.00	-0.01	0.01	0.06	0.00	0.21	0.14	0.10	-0.11	-0.08	-0.06	0.10	0.10	0.08	0.07	0.07	0.07	0.06
OBP (15games)	-0.01	0.01	0.00	0.00	0.09	0.00	0.17	0.19	0.14	-0.10	-0.10	-0.08	0.10	0.13	0.11	0.07	0.09	0.10	0.05
OBP (30games)	-0.01	0.01	0.00	0.01	0.12	0.00	0.12	0.16	0.17	-0.07	-0.09	-0.09	0.08	0.13	0.15	0.05	0.09	0.11	0.04
Temperature	-0.01	-0.01	0.00	0.00	0.06	-0.01	0.03	0.04	0.04	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Humidity	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Altitude	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ESPN Hit Factor	-0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	-0.02	0.00	0.01	-0.01	-0.02	-0.03	0.00	0.00	0.01	0.00
1st Pixel	0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1st Open	-0.02	0.01	0.01	-0.01	0.00	0.00	0.01	0.01	0.01	-0.02	-0.02	-0.02	0.00	-0.01	-0.01	0.00	0.00	0.00	-0.01
1st Retreatable	0.02	-0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1st/A	-0.30	-0.41	0.00	0.00	0.05	-0.02	0.02	0.02	0.02	-0.02	-0.02	-0.02	0.02	0.04	0.04	0.00	0.00	0.01	0.01
1st/H	-0.44	-0.55	-0.01	0.00	0.07	-0.02	0.02	0.02	0.03	-0.02	-0.02	-0.02	0.02	0.04	0.05	0.00	0.00	0.01	0.01
1st/R	-0.44	-0.35	-0.01	0.00	0.01	-0.02	0.02	0.03	0.03	-0.02	-0.02	-0.03	0.02	0.04	0.05	0.00	0.00	0.01	0.01
1st/R/A	0.64	0.48	0.01	0.01	0.01	0.09	0.03	0.05	0.06	0.02	0.03	0.03	-0.01	-0.01	-0.02	0.01	0.01	0.01	0.03
Hit (2)	0.00	0.01	0.00	0.01	0.02	-0.08	0.03	0.04	0.05	-0.04	-0.05	-0.05	0.02	0.02	0.03	0.01	0.01	0.01	0.00

Variable	28% (15games)	28% (30games)	HR% (7games)
Batter Hand (2)	0.01	0.00	0.03
Throwing Hand	0.00	0.00	0.00
Road/Home	0.00	0.00	0.00
HRs vs Pitcher	0.01	0.04	0.00
Batter game	0.00	0.00	0.00
Batting Order	0.01	0.04	0.00
HRs (7games)	-0.06	-0.12	-0.05
HRs (15games)	-0.09	-0.09	-0.05
HRs (30games)	-0.11	-0.12	-0.07
50% (7games)	-0.09	-0.09	-0.07
50% (15games)	-0.11	-0.11	-0.08
50% (30games)	-0.09	-0.09	-0.07
BB% (7games)	0.11	0.11	0.08
BB% (15games)	0.17	0.17	0.12
BB% (30games)	0.12	0.12	0.09
28% (7games)	0.68	0.74	0.01
28% (15games)		0.74	0.01
28% (30games)	0.74		0.01
HR% (7games)	0.01	0.01	
HR% (15games)	0.01	0.01	0.72
HR% (30games)	0.01	0.02	0.59
AB (7games)	0.01	0.01	0.01
AB (15games)	0.01	0.01	0.00
AB (30games)	0.01	0.01	0.01
Hit Streak	0.00	0.00	0.00
Avg Exit Velocity	0.00	0.00	0.00
Avg Launch Angle	0.00	0.00	0.00
Barrels/PA %	0.00	0.00	0.00
Percentage Shift	0.02	0.02	0.01
OBP (7games)	0.05	0.07	0.06
OBP (15games)	0.10	0.07	0.05
OBP (30games)	0.09	0.09	0.04
Temperature	0.00	0.00	0.00
Humidity	0.00	0.00	0.00
Altitude	0.00	0.00	0.00
ESPN Hit Factor	0.01	0.02	0.00
1st Pixel	0.00	0.00	0.01
1st Open	0.00	0.00	-0.01
1st Retreatable	0.00	0.00	0.00
1st/A	0.01	0.01	0.01
1st/H	0.01	0.01	0.01
1st/R	0.01	0.01	0.01
1st/R/A	0.01	0.01	0.03
Hit (2)	0.01	0.02	0.00

Variables	HR% (15games)	HR% (30games)	AB (7games)	AB (15games)	AB (30games)	Hit Streak	Average Launch Angle	Average Exit Velocity	Brls/PA %	Percentage Shift	OBP (7games)	OBP (15games)	OBP (30games)	Pitcher Game	S Hit/Inn (3 games)	S Hit/Inn (5 games)	S Hit/Inn (10games)	B Hit/Inn (3 games)	B Hit/Inn (5 games)	B Hit/Inn (10 games)
Batter Hand (2)	0.04	0.05	0.00	0.00	0.01	0.00	0.02	0.06	0.05	-0.40	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01
Throwing Hand	0.00	0.00	0.01	0.02	0.01	0.01	-0.01	0.00	0.00	0.11	0.00	0.01	0.01	-0.02	0.01	0.01	0.01	0.00	0.00	0.01
Road/Home	0.00	0.00	-0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.01	0.01	0.01	-0.03	-0.01	0.01
H% vs Pitcher	0.00	0.00	0.03	0.04	0.04	0.01	-0.02	0.00	-0.01	-0.03	0.01	0.00	0.01	0.00	0.02	0.03	0.03	0.00	0.00	0.00
batter game	0.12	0.18	0.19	0.30	0.44	0.03	0.17	0.15	0.07	0.18	0.06	0.09	0.12	0.38	-0.01	-0.01	0.00	0.00	0.00	0.00
Batting Order	-0.08	-0.13	-0.42	-0.47	-0.47	-0.13	-0.01	-0.21	-0.13	-0.08	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
H% (7games)	0.13	0.07	0.27	0.21	0.16	0.45	-0.08	0.05	0.00	-0.08	0.21	0.17	0.12	-0.02	0.00	0.01	0.00	0.05	0.04	0.04
H% (15games)	0.18	0.10	0.29	0.31	0.23	0.33	-0.12	0.06	0.00	-0.11	0.14	0.19	0.16	-0.03	0.00	0.00	0.00	0.03	0.03	0.03
H% (30games)	0.10	0.13	0.31	0.35	0.33	0.26	-0.15	0.07	-0.01	-0.14	0.10	0.14	0.17	-0.03	0.00	0.00	0.00	0.03	0.02	0.02
SO% (7games)	0.13	0.16	-0.09	-0.13	-0.15	-0.15	0.12	0.12	0.22	0.15	-0.11	-0.10	-0.07	0.04	0.00	0.00	0.00	-0.02	-0.02	-0.02
SO% (15games)	0.15	0.20	-0.14	-0.16	-0.18	-0.12	0.15	0.15	0.27	0.18	-0.08	-0.10	-0.09	0.05	0.00	0.00	0.00	-0.02	-0.01	-0.02
SO% (30games)	0.19	0.22	-0.17	-0.20	-0.21	-0.13	0.16	0.17	0.30	0.21	-0.06	-0.08	-0.09	0.06	0.00	0.00	0.00	-0.01	-0.01	-0.01
BB% (7games)	0.09	0.07	-0.01	-0.02	0.01	0.20	0.01	0.09	0.06	0.03	0.10	0.10	0.08	0.00	0.00	0.00	0.01	0.03	0.02	0.02
BB% (15games)	0.14	0.11	0.03	0.02	0.03	0.16	0.01	0.13	0.08	0.04	0.10	0.13	0.13	0.00	0.00	0.00	0.00	0.02	0.01	0.01
BB% (30games)	0.11	0.14	0.07	0.07	0.07	0.13	0.02	0.16	0.10	0.05	0.08	0.11	0.15	0.00	0.00	0.00	0.00	0.02	0.01	0.01
2B% (7games)	-0.02	-0.01	0.04	0.02	0.00	0.16	0.04	0.04	0.03	0.01	0.07	0.07	0.05	0.00	0.00	0.00	0.00	0.01	0.01	0.01
2B% (15games)	-0.01	0.00	0.09	0.09	0.06	0.12	0.06	0.08	0.05	0.02	0.07	0.10	0.09	0.00	0.00	0.01	0.00	0.01	0.01	0.01
2B% (30games)	0.02	0.02	0.13	0.15	0.14	0.11	0.08	0.11	0.07	0.04	0.06	0.09	0.11	0.00	0.00	0.01	0.00	0.02	0.01	0.01
HR% (7games)	0.72	0.59	0.01	0.00	0.01	0.11	0.23	0.26	0.33	0.24	0.06	0.05	0.05	0.03	0.00	0.01	0.01	0.01	0.01	0.01
HR% (15games)	[redacted]	0.81	0.03	0.03	0.03	0.06	0.32	0.36	0.44	0.33	0.05	0.07	0.06	0.04	0.00	0.00	0.00	0.01	0.01	0.00
HR% (30games)	0.81	[redacted]	0.06	0.08	0.09	0.05	0.39	0.45	0.55	0.42	0.05	0.07	0.08	0.06	0.00	0.00	0.01	0.01	0.01	0.00
AB (7games)	0.03	0.06	[redacted]	0.77	0.61	0.18	-0.05	0.08	0.02	-0.03	0.13	0.11	0.08	0.00	0.00	0.00	0.00	0.04	0.03	0.03
AB (15games)	0.03	0.08	0.77	[redacted]	0.82	0.14	-0.04	0.10	0.03	-0.03	0.07	0.11	0.10	0.04	0.01	0.00	0.00	0.02	0.02	0.02
AB (30games)	0.03	0.09	0.61	0.82	[redacted]	0.12	-0.03	0.12	0.03	-0.01	0.05	0.07	0.10	0.10	0.00	0.00	0.00	0.01	0.01	0.02
Hit Streak	0.06	0.05	0.18	0.14	0.12	[redacted]	-0.03	0.04	0.01	0.03	0.14	0.10	0.07	0.02	0.00	0.00	0.00	0.06	0.05	0.04
Average Launch Angle	0.32	0.39	-0.05	-0.04	-0.03	-0.03	[redacted]	0.20	0.39	0.51	0.01	0.01	0.02	0.07	0.00	0.00	0.00	0.00	0.00	0.00
Average Exit Velocity	0.36	0.45	0.08	0.10	0.12	0.04	0.20	[redacted]	0.66	0.40	0.03	0.03	0.04	-0.02	0.00	-0.01	-0.01	0.00	0.00	0.00
Brls/PA %	0.44	0.55	0.02	0.03	0.03	0.01	0.39	0.66	[redacted]	0.49	0.02	0.02	0.02	-0.05	0.00	0.00	0.00	0.00	0.00	0.00
Percentage Shift	0.33	0.42	0.03	-0.03	-0.01	-0.03	0.51	0.40	0.49	[redacted]	0.01	0.01	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00
OBP (7games)	0.05	0.05	0.13	0.07	0.05	0.14	0.01	0.03	0.02	0.01	[redacted]	0.68	0.52	0.02	0.01	0.01	0.02	0.15	0.12	0.11
OBP (15games)	0.07	0.07	0.11	0.11	0.07	0.10	0.01	0.03	0.02	0.01	0.68	[redacted]	0.74	0.03	0.02	0.02	0.03	0.10	0.08	0.09
OBP (30games)	0.06	0.08	0.08	0.10	0.10	0.07	0.02	0.04	0.02	0.02	0.52	0.74	[redacted]	0.06	0.02	0.03	0.03	0.08	0.07	0.07
Pitcher Game	0.04	0.06	0.00	0.04	0.10	-0.02	0.07	-0.02	-0.05	0.03	0.02	0.03	0.06	[redacted]	-0.07	-0.08	-0.09	-0.02	-0.02	-0.03
S Hit/Inn (3 games)	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.02	0.02	-0.07	[redacted]	0.81	0.66	0.02	0.03	0.06
S Hit/Inn (5 games)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.01	0.02	0.03	-0.08	0.81	[redacted]	0.80	0.02	0.04	0.07
S Hit/Inn (10 games)	0.00	0.01	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.02	0.03	0.03	-0.09	0.66	0.80	[redacted]	0.03	0.04	0.07
B Hit/Inn (3 games)	0.01	0.01	0.04	0.02	0.01	0.06	0.00	0.00	0.00	0.00	0.15	0.10	0.08	-0.02	0.02	0.02	0.03	[redacted]	0.76	0.53
B Hit/Inn (5 games)	0.01	0.01	0.03	0.02	0.01	0.05	0.00	0.00	0.00	0.00	0.12	0.08	0.07	-0.02	0.03	0.04	0.04	0.76	[redacted]	0.70
B Hit/Inn (10 games)	0.00	0.00	0.03	0.02	0.02	0.04	0.00	0.00	0.00	0.00	0.11	0.09	0.07	-0.03	0.06	0.07	0.07	0.53	0.70	[redacted]
Temperature	0.02	0.02	0.03	0.05	0.07	0.02	-0.02	0.00	0.00	-0.02	0.03	0.05	0.04	0.04	0.02	0.02	0.00	0.04	0.04	0.06
WindSpd	-0.02	-0.02	0.00	0.00	-0.01	0.01	0.02	-0.03	-0.03	0.00	-0.01	0.00	0.00	0.00	0.02	0.02	0.03	0.00	0.01	0.01
Altitude	-0.02	-0.03	0.01	0.01	0.01	0.02	-0.02	-0.02	-0.02	-0.02	0.03	0.02	0.04	0.02	0.00	0.01	0.02	0.05	0.04	0.04
ESPN Hit Factor	0.00	0.00	0.02	0.02	0.02	0.03	-0.01	-0.02	-0.02	-0.01	0.04	0.03	0.05	0.06	0.02	0.02	0.03	0.05	0.04	0.02
S Fixed	0.02	0.01	-0.01	-0.01	-0.01	-0.01	0.02	-0.01	0.00	0.02	0.02	0.00	-0.01	-0.01	-0.01	-0.02	-0.03	0.00	-0.01	0.00
S Open	-0.02	-0.02	0.00	0.00	0.00	0.01	0.03	-0.03	-0.02	0.00	0.01	0.01	0.01	0.00	0.02	0.03	0.02	0.02	0.03	0.03
S Retractable	0.01	0.02	0.00	0.00	0.00	-0.01	-0.04	0.03	0.02	-0.01	-0.01	-0.01	0.00	0.01	-0.01	-0.02	-0.01	-0.02	-0.03	-0.03
S L/L	0.01	0.02	0.03	0.04	0.04	0.02	0.00	0.02	0.02	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S L/R	-0.05	-0.06	-0.02	-0.02	-0.03	-0.01	-0.02	-0.08	-0.06	0.34	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.01
S R/L	0.00	0.00	-0.03	-0.04	-0.04	-0.02	0.01	-0.01	-0.01	-0.20	-0.01	-0.01	-0.01	0.02	-0.01	-0.01	-0.01	0.00	0.00	-0.01
S R/R	0.04	0.05	0.03	0.04	0.04	0.02	0.01	0.08	0.06	-0.24	0.00	0.00	0.00	-0.02	0.01	0.00	0.00	0.00	0.00	0.00
Hit (2)	-0.01	0.00	0.05	0.06	0.06	0.02	-0.02	0.03	0.01	-0.02	0.01	0.01	0.01	-0.02	0.02	0.02	0.02	0.02	0.01	0.01

Variables	Temperature	WindSpd	Altitude	ESPN Hit Factor	Is Fixed	Is Open	Is Retractable	Is L/L	Is L/R	Is R/L	Is R/R	Hit (2)
Batter Hand (2)	-0,01	-0,01	-0,01	-0,01	0,01	-0,02	0,02	-0,30	-0,88	0,44	0,64	0,00
Throwing Hand	0,01	-0,01	0,02	0,00	0,00	0,01	-0,01	-0,41	0,45	-0,85	0,48	0,01
Road/Home	0,00	0,00	0,00	0,00	-0,01	0,01	0,00	0,00	0,00	-0,01	0,01	0,00
H% vs Pitcher	0,00	0,00	0,00	0,00	-0,01	0,00	0,00	0,00	-0,01	0,00	0,01	0,01
Batter game	0,06	-0,01	0,02	0,03	-0,01	-0,01	0,01	0,05	-0,01	-0,03	0,01	0,02
Batting Order	-0,01	0,00	-0,01	0,00	0,01	0,00	0,00	-0,02	-0,09	0,01	0,09	-0,08
H% (7games)	0,03	0,00	0,02	0,01	0,00	0,01	-0,01	0,02	-0,03	-0,02	0,03	0,03
H% (15games)	0,04	-0,01	0,02	0,02	-0,01	0,01	0,00	0,02	-0,04	-0,02	0,05	0,04
H% (30games)	0,04	-0,01	0,02	0,02	-0,01	0,01	0,00	0,02	-0,05	-0,02	0,06	0,05
SO% (7games)	-0,01	-0,01	0,00	0,00	0,03	-0,02	0,01	-0,02	-0,03	0,03	0,02	-0,04
SO% (15games)	0,00	-0,01	0,00	0,00	0,03	-0,02	0,01	-0,02	-0,04	0,03	0,03	-0,05
SO% (30games)	-0,01	-0,01	0,00	0,01	0,04	-0,02	0,01	-0,02	-0,04	0,03	0,03	-0,05
BB% (7games)	0,00	-0,02	0,01	-0,01	0,00	-0,01	0,01	0,02	0,02	-0,02	-0,01	0,02
BB% (15games)	0,00	-0,02	0,01	-0,02	0,00	-0,02	0,02	0,04	0,02	-0,04	-0,01	0,02
BB% (30games)	0,00	-0,01	0,01	-0,03	-0,01	-0,01	0,02	0,04	0,04	-0,05	-0,02	0,03
2B% (7games)	0,00	-0,01	0,01	0,00	0,00	0,00	0,00	0,00	-0,01	0,00	0,01	0,01
2B% (15games)	0,00	-0,01	0,02	0,01	0,00	0,00	0,00	0,01	-0,01	-0,01	0,01	0,01
2B% (30games)	0,00	0,00	0,02	0,02	-0,01	0,00	0,00	0,01	-0,01	-0,01	0,01	0,02
HR% (7games)	0,01	-0,01	-0,01	0,00	0,01	-0,01	0,01	0,01	-0,04	0,00	0,03	0,00
HR% (15games)	0,02	-0,02	-0,02	0,00	0,02	-0,02	0,01	0,01	-0,05	0,00	0,04	-0,01
HR% (30games)	0,02	-0,02	-0,03	0,00	0,01	-0,02	0,02	0,02	-0,06	0,00	0,05	0,00
AB (7games)	0,03	0,00	0,01	0,02	-0,01	0,00	0,00	0,03	-0,02	-0,03	0,03	0,05
AB (15games)	0,05	0,00	0,01	0,02	-0,01	0,00	0,00	0,04	-0,02	-0,04	0,04	0,06
AB (30games)	0,07	-0,01	0,01	0,02	-0,01	0,00	0,00	0,04	-0,03	-0,04	0,04	0,06
Hit Streak	0,02	0,01	0,02	0,03	-0,01	0,01	-0,01	0,02	-0,01	-0,02	0,02	0,02
Average Launch Angle	-0,02	0,02	-0,02	-0,01	0,02	0,03	-0,04	0,00	-0,02	0,01	0,01	-0,02
Average Exit Velocity	0,00	-0,03	-0,02	-0,02	-0,01	-0,03	0,03	0,02	-0,08	-0,01	0,08	0,03
Brls/PA %	0,00	-0,03	-0,02	-0,02	0,00	-0,02	0,02	0,02	-0,06	-0,01	0,06	0,01
Percentage Shift	-0,02	0,00	-0,02	-0,01	0,02	0,00	-0,01	0,14	0,34	-0,20	-0,24	-0,02
OBP (7games)	0,03	-0,01	0,03	0,04	0,02	0,01	-0,01	0,00	0,01	-0,01	0,00	0,01
OBP (15games)	0,05	0,00	0,02	0,03	0,00	0,01	-0,01	0,00	0,01	-0,01	0,00	0,01
OBP (30games)	0,04	0,00	0,04	0,05	-0,01	0,01	0,00	0,00	0,01	-0,01	0,00	0,01
Pitcher Game	0,04	0,00	0,02	0,06	-0,01	0,00	0,01	0,00	0,00	0,02	-0,02	-0,02
S Hit/Inn (3 games)	0,02	0,02	0,00	0,02	-0,01	0,02	-0,01	0,00	0,00	-0,01	0,01	0,02
S Hit/Inn (5 games)	0,02	0,02	0,01	0,02	-0,02	0,03	-0,02	0,00	0,01	-0,01	0,00	0,02
S Hit/Inn (10 games)	0,00	0,03	0,02	0,03	-0,03	0,02	-0,01	0,00	0,00	-0,01	0,00	0,02
B Hit/Inn (3 games)	0,04	0,00	0,05	0,05	0,00	0,02	-0,02	0,00	0,00	0,00	0,00	0,02
B Hit/Inn (5 games)	0,04	0,01	0,04	0,04	-0,01	0,03	-0,03	0,00	0,01	0,00	0,00	0,01
B Hit/Inn (10 games)	0,06	0,01	0,04	0,02	0,00	0,03	-0,03	0,00	0,01	-0,01	0,00	0,01
Temperature		-0,07	0,21	0,16	-0,07	0,06	-0,03	-0,01	0,01	-0,01	0,00	0,02
WindSpd	-0,07		-0,03	0,19	-0,27	0,52	-0,43	0,00	0,01	0,00	-0,02	0,01
Altitude	0,21	-0,03		0,39	-0,05	0,09	-0,07	-0,01	0,01	-0,02	0,01	0,02
ESPN Hit Factor	0,16	0,19	0,39		-0,23	0,30	-0,21	0,01	0,01	-0,01	-0,01	0,04
Is Fixed	-0,07	-0,27	-0,05	-0,23		-0,34	-0,09	0,00	0,00	0,00	0,01	-0,02
Is Open	0,06	0,52	0,09	0,30	-0,34		-0,91	0,00	0,02	-0,01	-0,02	0,01
Is Retractable	-0,03	-0,43	-0,07	-0,21	-0,09	-0,91		0,00	-0,02	0,01	0,02	-0,01
Is L/L	-0,01	0,00	-0,01	0,01	0,00	0,00	0,00		-0,19	-0,13	-0,19	0,00
Is L/R	0,01	0,01	0,01	0,01	0,00	0,02	-0,02	-0,19		-0,39	-0,56	0,00
Is R/L	-0,01	0,00	-0,02	-0,01	0,00	-0,01	0,01	-0,13	-0,39		-0,41	0,00
Is R/R	0,00	-0,02	0,01	-0,01	0,01	-0,02	0,02	-0,19	-0,56	-0,41		0,01
Hit (2)	0,02	0,01	0,02	0,04	-0,02	0,01	-0,01	0,00	0,00	0,00	0,01